A COX-BASED RISK PREDICTION MODEL FOR EARLY DETECTION OF CARDIOVASCULAR DISEASE

A THESIS SUBMITTED TO AUCKLAND UNIVERSITY OF TECHNOLOGY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF COMPUTER AND INFORMATION SCIENCES

Supervisor

Dr. Farhaan Mirza

Associate Prof. Hamid GholamHosseini

Dr. Mirza Baig

September 2018

By

Xiaona Jia

School of Engineering, Computer and Mathematical Sciences

Abstract

Cardiovascular disease (CVD) is the number one cause of mortality around the world. A fair proportion of health care resource is consumed for managing CVD, which imposes a heavy health burden on the community. To prevent the prevalence of CVD, an effective approach is to create prediction models to assess the CVD risk and then enable early lifestyle adjustments or clinical treatments. A great amount of research has been done but challenges and issues still exist. The aim of this research is to create an effective risk prediction model for early assessment and detection of CVD.

The Framingham Original Cohort Study data set of 5079 subjects aged from 30 to 74 years old who had not previous symptoms of CVD at the baseline was enclosed. The Cox regression method was used for the data analysis. A complete process of creating risk models was conducted according to statistical regression strategies. Lastly, a risk prediction model for general CVDs was generated based on risk predictors, including age, sex, body mass index, hypertension, pulse rate, systolic blood pressure, cigarettes per day, and diabetes. We obtained a good predictive ability of discrimination and calibration with ROC of 0.71 indicating a good accuracy for the risk estimate of CVD.

In our new Cox-based risk model, a novel predictor heart rate was incorporated to predict CVD risk, which expands the predictive ability of existing CVD risk models. Moreover, this risk prediction model was developed based on office risk factors, i.e. the measure of risk factors does not require clinical tests, which would be beneficial to both health care providers and patients to assess CVD event rates at any time and any place.

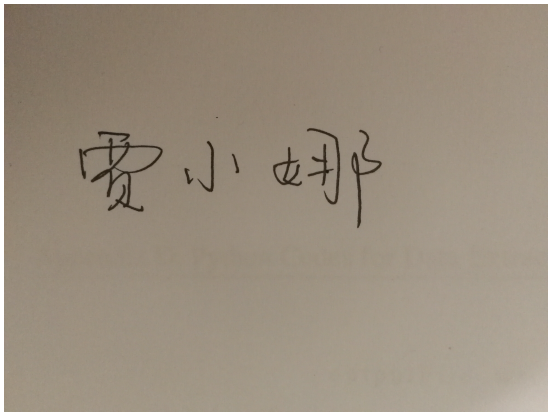
Contents

List of Tables

List of Figures

Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the qualification of any other degree or diploma of a university or other institution of higher learning.



Signature of student

Acknowledgements

Many thanks to:

My supervisory team, Dr. Farhaan Mirza, Associate Prof. Hamid GholamHosseini, and Dr. Mirza Baig, School of Engineering, Computer and Mathematical Sciences, Auckland University of Technology. This research area is totally new for me and I did not have much confidence at the beginning but they always give me encouragement and constructive suggestions, and steer me in the right research direction. Whenever I ran into difficulties, their offices are always open to me. I'm appreciated the help they provide to me. They not only guide me to complete my academic research thesis but also let me to believe that nothing is impossible.

My husband who supports me throughout my challenging but exciting master journey. Whenever I am encountered either research troubles or life difficulties, he is always there giving me spiritual support and encouragement. I'd love to say "you are my soul mate, and forever". Thank a lot my parents-in-law, they sacrifice so much their free time to help me take care of my little daughter. Also, I will never forget my little girl supporting her mummy at the corner of the earth thousands of miles away. Without the love and encouragement of them, I would not go so far.

My friends, Ling Liang, Xiaoxi Ruan, Yu liu. They give me sincere support and recommendation when I ask for advice from them. Also, their companionship is precious for me studying in a foreign country. Finally, I appreciated the help from the team in Auckland University of Technology Ethics Committee (AUTEC).

Chapter 1

Introduction

1.1 Research Background

Cardiovascular disease (CVD) describes a variety of types of conditions relevant to the heart and blood system. It includes stroke, rheumatic heart disease, Coronary Artery Diseases (CAD), heart attack, Heart Failure (HF), Coronary Heart Disease (CHD), etc (Mendis, Puska, Norrving et al., 2011). The prevalence of CVD has increased in recent decades. The 2015 statistics on heart disease in America indicate that the percentage of people aged 20 to 40 years old who have CVD was 11%, 37% for people aged 40 to 60 years, 71% for people aged 60 to 80 years, and 85% for people over 80 years old (Mozaffarian et al., 2015). In New Zealand, a high proportion of the population who have not developed CVD has an over 5% CVD risk in the next few years. The Ministry of Health of New Zealand says that the percentage of population who have a risk of CVD higher than 5% is 26% (Ministry of Health, 2018).

Because of the high incidence of disease, CVD has become the leading cause of mortality around the world. In 2003, CVD was the number one cause of death in New Zealand (Chan et al., 2008). Statistics on CVD mortality in 2008 suggest that the percentage of deaths caused by CVD is 45% for females and 43% for males (Statistics

New Zealand, 2012). Thanks to effective prevention and treatment efforts, the mortality rates caused by CVD have decreased in recent years. The percentage of deaths due to CVD has declined to 33% in 2017 but there were still 172,000 people living with heart disease (Heart Foundation, 2017). CVD remains the second most common cause of death in New Zealand (Statistics New Zealand and Ministry of Pacific Island Affairs, 2011). Because of the high rate of disease incidence and death caused by CVD, a high proportion of resource has been consumed by health care for CVD, which is imposing a heavy health burden on the community. For example, in the United States, 17% of overall national health care expenditure is consumed by the treatment of CVD (Heidenreich et al., 2011).

Many cardiovascular-related deaths are premature and preventable. McGill, McMahan and Gidding (2008) think that the health condition and disease symptoms can be improved by effective health management such as the effective dietary and lifestyle interventions, and drug intervention. In fact, 90% of CVD is preventable (McGill et al., 2008). In order to achieve the prevention of CVD, an effective approach is to assess and record the risk of CVD before it happens and then enable early clinical treatments or lifestyle adjustments.

A significant number of studies have been done on the risk estimation of CVD in the past decades (Damen et al., 2016). CVD risk prediction models have proved to be effective in the health and disease management for clinicians and patients. The new PREDICT CVD risk assessment equation developed for primary health care among the population in New Zealand has been integrated into the general practice electronic health records (EHRs) and a web-based software called 'PREDICT' (Wells et al., 2017). This 'PREDICT' web platform has got 400,728 patients assessed for CVD risk and is becoming an effective tool for decision support and a management system for general practitioners. Another risk estimation tool, the old Framingham CVD risk prediction equation developed by Wilson et al. (1998), has been incorporated into the New Zealand

Primary Care Handbook (Cardiovascular Disease Risk Assessment Steering Group and others, 2017). This handbook guideline is a convenient ready-reference for primary care practitioners and has helped countless New Zealanders to assess their CVD health conditions.

Consequently, articles of relevant prediction models increase with years. There was an obvious peak in the last ten years (Damen et al., 2016). A series of CVD risk assessment equations have been developed, such as the Framingham scores from the Framingham Heart Study (FHS) (D'Agostino et al., 2008; Lloyd-Jones et al., 2004), the QRISK equations (Hippisley-Cox et al., 2007), the Europe SCORE risk models (Conroy et al., 2003), the ASSIGN scores from the Scottish Heart Health Extended Cohort (SHHEC) (Woodward, Brindle & Tunstall-Pedoe, 2006), the Prospective Cardiovascular Münster(PROCAM) equations (Assmann, Cullen & Schulte, 2002), and the CUORE Cohort Study formulas (Ferrario et al., 2005).

1.2 Research Motivation

Risk prediction for CVD requires a two-stage procedure. The first stage involves the measurement of observations and complete prospective biological information. The collection of longitudinal data is a process of long-time follow up, normally taking years to finish the full project, which is difficult in large cohorts. Some long-term prospective studies (D'Agostino et al., 2008; Woodward et al., 2006; Ferrario et al., 2005; Hippisley-Cox et al., 2007) not only derived risk prediction equations but also provided valuable designs, work-flow guidance of data collection, as well as data resources for later researchers.

The performance of most of the CVD prediction models reviewed above has been internally or externally validated to make sure they have an accurate ability of predicting the probability of developing CVD (D'Agostino Sr et al., 2001; D'Agostino et al., 2008;

Hippisley-Cox et al., 2007; Hippisley-Cox, Coupland, Vinogradova, Robson & Brindle, 2008). Some have been applied to the general practice (D'Agostino et al., 2008; Assmann et al., 2002). However, challenges and issues in the development of CVD risk estimation models still exist:

- Risk models using a single factor for the risk estimation of CVD cannot see the influence of multiple factors simultaneously. The Kailuan model devised by Yu et al. (2017) on the basis of analysing the cumulative exposure to Resting Heart Rate (cumRHR), the 10-year Atherosclerotic Cardiovascular Disease (ASCVD) model developed by Han, Park, Kim, Kim and Chun (2017), and the Sudden Cardiac Arrest (SCA) model developed by Murukesan, Murugappan, Iqbal and Saravanan (2014) are all single-factor-based risk models for one specific individual component of CVDs.

- Existing CVD risk prediction models using statistical regression analysis (such as the Weibull (Cannon, 2012), the Kaplan-Meier (Kaplan & Meier, 1958), or the Cox regression analysis (Cox, 1992)) prefer to use classic predictors such as sex, age, smoking, diabetes, high blood pressure, and total cholesterol, etc., to estimate the risk score. The new version of the Framingham laboratory-based risk model was generated based on seven conventional risk factors: age, Systolic Blood Pressure (SBP), anti-hypertensive medication treatment for hypertension, total cholesterol, High Density Lipoprotein (HDL) cholesterol, current smoking, and diabetes status (D'Agostino et al., 2008). Other similar risk models like the one devised by Unnikrishnan et al. (2016) and the QRISK (Hippisley-Cox et al., 2007) all use the same conventional predictors as the Framingham equation (D'Agostino et al., 2008).

- Several CVD risk assessment tools based on data mining or machine learning methods have novel predictors incorporated in the risk models but either a result

of classification and clustering (Hachesu, Ahmadi, Alizadeh & Sadoughi, 2013; J. Kim, Lee & Lee, 2015; Kumari & Godara, 2011; Melillo et al., 2015), a risk level of developing CVD (Vaanathi, 2017), or an ultimate risk value (Unnikrishnan et al., 2016; Murukesan et al., 2014), was given. An absolute risk estimation for how many years in the following time cannot be obtained, which is a limitation to supporting health management for decision making.

- A CVD risk prediction equation cannot demonstrate its value unless it is applied to general practice. These risk estimation formulas have been widely applied to clinical guidelines and web-based software tools. The New Zealand Primary Care Handbook incorporated the old version of the Framingham CVD prediction algorithms (Wilson et al., 1998) for New Zealanders' primary health care management. The web-based platform 'PREDICT' integrated the CVD risk assessment equation developed by Wells et al. (2017) for estimating the patient's CVD and diabetes profile. However, CVD risk prediction models have not been applied to wearable monitoring systems for health care management. The lack of valid risk prediction algorithms for implementing smart wearable systems in real life is a great challenge for medical decision making (Lymberis, 2003).

1.3 Research Objectives

Motivated by the challenges and issues that current researchers are facing, we were encouraged to devise an effective prognosis platform for early detection of CVD. The aim of this research is to create a risk prediction model for early detection of CVD.

The scope of this research is defined as follows:

- Identifying risk factors that have effects on the development of CVD.

- Incorporating identified novel risk factors as well as conventional risk predictors into the CVD risk prediction model.
- Developing a multiple-variable-based risk prediction model targeting general CVD events. This model should have the ability to compute absolute risk scores in the next 10 years.
- Validating the fitted CVD risk prediction model to see its predictive performance.

A full process of conducting this research is summarised in Figure 1.1.

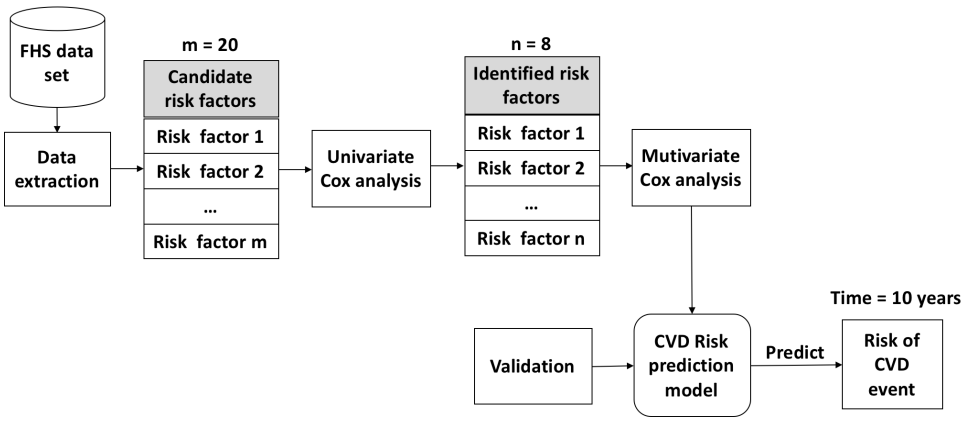


Figure 1.1: The Full Process of Conducting this Research
Notes: m = The number of candidate risk factors; n = The number of identified risk factors

1.4 Research Contributions

This thesis will make several contributions:

- We explore a novel risk factor (pulse rate) as a significant predictor affecting on the development of CVD.

- We generate a risk prediction model aimed at general CVDs. It is an office-based tool, i.e. practitioners can use it during an office visit without requiring a clinical laboratory test.
- An absolute CVD risk probability in 10 years can be obtained for an individual using three forms of risk estimation: risk equations, nomograms, and survival curves.

1.5 Research Benefits

The findings of this research will be beneficial to both health care providers and potential CVD patients.

- For primary health organisations: they could incorporate the CVD risk prediction models into clinical guidelines such as the health care hand book for CVD risk assessment and management.
- For medical physicians: they could use this risk prediction to calculate risk scores during a patient visit directly, so they can give recommendations of treatments based on adequate evidence, which will be particularly beneficial for persons with high CVD risk.
- For general individuals: this office-based general risk estimation model enables them to monitor their CVD risk trend at home, which increases the autonomy and involvement of patients regarding to their health care. Moreover, it is also a way to augment the relationships between physicians and patients, which revolutionises health care management.

1.6 Thesis Structure

This thesis is organised as follows:

Chapter 1 - Introduction: introduces the background of this research, including the challenges and issues of CVD risk prediction, research objectives, research contributions, and research benefits.

Chapter 2 - Literature review: elaborates on previous works on CVD detection in terms of current CVD risk prediction models, modelling methods, risk factors, research data sets, and eHealth Solutions for CVD Prediction.

Chapter 3 - Research method: discusses the design of the research method and procedures of implementing this research.

Chapter 4 - Implementation of the risk estimation model: states the whole process of developing a CVD risk prediction model using Cox regression analysis guided by the regression modelling strategies.

Chapter 5 - Validation: validates the accuracy of the fitted CVD risk prediction model in terms of statistical perspectives (discrimination and calibration), as well as empirical perspectives.

Chapter 6 - Findings and discussion: presents the findings of this research and discusses the contributions and implications with respect to the findings.

Chapter 7 - Conclusion: summarises the main work and the findings of this research and illustrates limitations and potential future works.

Chapter 2

Literature Review

2.1 Introduction

In this chapter, we provide a review of current literature on CVD risk prediction as well as CVD prediction applications. Articles and journals will be researched section by section.

Section 2.2 presents an overview of the CVD outcomes that current researchers are targeting.

Section 2.3 introduces the popular modelling methods applied to develop a risk prediction model, and risk factors that current CVD prediction models have included.

Section 2.4 reviews the available data sets for CVD research in terms of the data features, study populations, risk factors and sample size.

Section 2.5 gives the existing literature on eHealth solutions for CVD prediction, including clinical guidelines, web-based tools, and mHealth applications.

2.2 CVD Overview

CVD outcomes are events or groups of events that a prediction model focuses on (Damen et al., 2016). It can be one specific component of CVD events such as CHD, CAD, stroke, etc., or a combination of these diseases, or the complete set of CVDs. In other words, CVD outcomes are the target diseases predicted from a risk model. The definition of the predicted outcome decides the inclusion of risk factors, the strategy of data collection, the selection of modelling algorithms, and consequently leads to different predicted models and different strategies of treatment (Damen et al., 2016). Thus, clearly defining the CVD outcome at the beginning of any research is critical.

Many kinds of CVD outcomes have been identified in the primary studies. Some prediction models target specific components of CVD. The 10-year risk prediction model developed by Wilson et al. (1998) defined CHD as the research target. Another risk equation developed by Lloyd-Jones et al. (2004) also targeted CHD but expanded the predictive time length to the lifetime. The 5-year risk estimation model (Butler et al., 2008) has the ability to predict Incident Heart Failure (IHF) among elders. These reviewed models have been fully validated and incorporated into guidelines of primary care (Expert Panel on Detection et al., 2001), which can be requested conveniently either online or in hard copy. However, they lack multiplicity. Only an individual CVD can be predicted, which limits their use in primary care.

In some cases, the risk assessment for a specific event is useful when physicians would like to assess and treat a particular component of CVDs (Jackson, 2000). However, in primary care, physicians are mainly concerned with the prevention and health maintenance for general CVDs (Goff et al., 2014). They would like to know the probability of getting a specific CVD event using a general risk prediction tool. Because of this, some studies have started to focus on the risk prediction of general CVD (Hippisley-Cox et al., 2007; D'Agostino et al., 2008; Pencina, D'agostino, Larson,

Massaro & Vasan, 2009). Researchers from the Framingham Heart Study (FHS) not only developed risk models for individual components of CVD events (Kannel, Feinleib, McNamara, Garrison & Castelli, 1979; Lloyd-Jones et al., 2004) but also published two risk functions for general CVD based on traditional risk factors and non-laboratory predictors (D'Agostino et al., 2008; Pencina et al., 2009). CVD risk scores can be estimated as the guidance of preventive care.

2.3 CVD Prediction

2.3.1 CVD Prediction Models

Along with the increase of CVD occurrence (Mozaffarian et al., 2015; Statistics New Zealand, 2012; Heart Foundation, 2017) and correspondingly increased medical cost (Heidenreich et al., 2011), the research on CVD prediction models is becoming popular. Risk prediction models have proven to be playing an important role in decision making in the clinical domain. It complements clinical reasoning for physicians to make treatment decisions (Moons et al., 2012). In personal health management, prediction models can assist individuals in the self-management of their health condition (Moons et al., 2012).

Research articles reporting CVD detection models have been annually increasing in recent decades. A systematic review of Clinical Prediction Models (CPMs) for CVD by Wessler et al. (2015) suggests that there were only three articles relevant to CVD risk prediction models in 1990 but the number had increased to approximately 500 by 2012. Figure 2.1 shows the increasing trend of published articles from 1990 to 2012.

There are two types of CVD prediction models: prognostic models and diagnostic models (Cui, 2009). Diagnostic models are used where a specific disease needs to be diagnosed. The physicians may not know the disease outcome of an individual. They

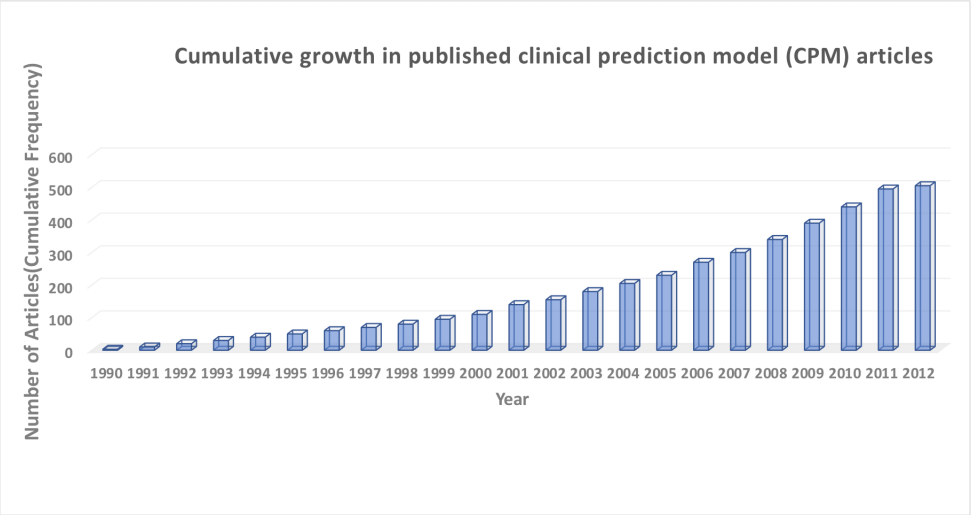


Figure 2.1: Articles Related to CVD Prediction Models from 1990 to 2012 (Wessler et al., 2015)

use a diagnostic model to assess the diagnostic factors for an individual and then a diagnosed result will be obtained. This examination result is only used for the treatment of this individual. The model developed by Mostafa et al. (2010) is a typical diagnostic model to diagnose Type 2 Diabetes Mellitus (T2DM).

Contrary to the diagnostic model, the nature of prognostic prediction models is that disease outcomes will not be determined after assessing the markers but a risk of developing these outcomes will be given For example, the FHS CHD risk model (Lloyd-Jones et al., 2004) introduced above can estimate the risk of having CHD in the next ten years. Prognostic models based on risk factors can be applied to new patients or individuals. New populations may be from different countries or different ethics (Moons, Altman, Vergouwe & Royston, 2009).

Based on the discussion on the nature of diagnostic models and prognostic models, we can see that they should be used in different settings and not be interchangeable. Diagnostic prediction models are commonly applied in clinical practice such as the hospital. Doctors could make a diagnosis based on a diagnostic prediction model. The prognostic models have a wider range of application areas like clinical guidelines,

self-management tools, and so on. Table 2.1 summarises the input, output, nature, and work environment of the prognostic model and the diagnostic model.

Table 2.1: Types of CVD Prediction Models

	Prognostic Model	Diagnostic Model
Input	Risk factors	Diagnostic factors
Output	Risk of developing a specific disease	Diagnosis of a specific disease
Nature	For risk estimation	For clinical diagnose
Work Environment	Hospital and similar practices	Clinical guideline, self-management applications, etc.

2.3.2 Modelling Methods

Numerous techniques can be utilised for generating CVD risk prediction models such as expert system (Bhatla & Jyoti, 2012), regression analysis (Kleinbaum, Kupper, Nizam & Rosenberg, 2013), fuzzy logic and rough set (J. Kim et al., 2015), data mining (Koh, Tan et al., 2011), etc. Regression analysis is the traditional method to be used, yet, along with the development of computer and information sciences, techniques like expert systems, classification and clustering have been applied to CVD detection in recent years.

2.3.2.1 Regression Analysis

Regression analysis is a traditional method for event prediction. A great number of methods are available such as the linear regression model, the binary logistic regression model, and the polynomial regression method (Kleinbaum et al., 2013). In the clinical field, the popular statistical method used for CVD prediction is survival analysis. Survival analysis (Despa, 2010) investigates the relation between the emergence of an event of interest and the expected duration of time it takes, such as cancer studies in biological organisms (Ma et al., 2008). Typical research questions in survival analysis are:

- What are significant characteristics that impact on the patient's survival?
- What is the risk for a person to have a defined event (disease)?
- What is the probability that an individual survives a certain duration of time?

Three types of statistical approaches can be used in survival analysis: non-parametric approaches, semi-parametric approaches, and parametric approaches (E. T. Lee & Wang, 2003).

Non-parametric methods are typically used to estimate the survival function from lifetime data. Popular techniques are the Kaplan-Meier estimator (Kaplan & Meier, 1958) and the log-rank test (Mantel, 1966; Cox, 1992). These two methods are univariate analysis. Survival curves can be plotted by comparing the survival distributions of two samples but only show the effect under one factor. The impact of any other factor is ignored. In addition, non-parametric methods only work well when the input variable is categorical but not useful for the quantitative variables such as age, weight, or waist. Weiner et al. (2004) developed a CVD risk estimation using the Kaplan-Meier and the log-rank analysis, in which Kaplan-Meier was used to obtain the survival times among people without Chronic Kidney Disease (CKD) against those having CKD, and the log-rank test was used for differentiating the differences among different groups.

Weibull and gamma regression models are popular parametric approaches in survival analysis (F. Harrell, 2013). Weibull model and gamma are both generalised from the exponential model (Miller Jr, 2011). These models can do both univariate analysis and multivariate analysis but they assume that the failure rate for two subjects over time is constant, i.e. the distribution of the hazard for two people does not change. The Weibull model is particularly useful when the samples are very small but if the assumption does not hold, the fitted model would be invalid (Cannon, 2012). The well-known Europe SCORE project (Conroy et al., 2003) used the Weibull method to develop a risk score

system for clinical management. The risk estimation equation developed by Anderson, Odell, Wilson and Kannel (1991) is also based on the Weibull model.

Cox proportional hazards regression analysis (Cox, 1992) is a semi-parametric method. It aims at investigating the importance of various risk factors simultaneously on the survival time of individuals through hazard function and works well for both quantitative and categorical variables. The QRISK model in the United Kingdom (Hippisley-Cox et al., 2007), the ASSIGN model from the SHHEC study (Woodward et al., 2006), the Framingham CVD risk model (D'Agostino et al., 2008), the Prospective Cardiovascular Münster (PROCAM) model (Assmann et al., 2002), and the prediction equation for coronary events in the CUORE cohort study (Ferrario et al., 2005), were all developed based on the Cox proportional hazards model.

It was reviewed that approximately half of the CVD risk models using regression analysis was Cox proportional hazards regression analysis (constituting nearly 44% of all the studies included), the proportion of other models (Weibull, logistic regression, etc.) used is 41%, and the remaining articles were not clear (Damen et al., 2016). This review indicates that the Cox proportional hazard regression analysis was the most popular method in survival analysis.

2.3.2.2 Data Mining Methods

The objective of data mining is to discover the unrevealed patterns and useful information from a data set (Jilani, Yasin, Yasin & Ardil, 2009). Data mining algorithms have been applied to a wide range of big data analytics areas such as business prediction for marketing, sales, or customer support (Berry & Linoff, 1997) for a long time. Over the past two decades, these techniques have been used in clinical data analytic and health care for decision making of policy-makers or clinical physicians (Koh et al., 2011). Thus, apart from the traditional statistical methods reviewed above, some researchers have recently proposed risk assessment models using classification algorithms such

as Support Vector Machines (SVM), Artificial Neural Network (ANN), fuzzy logic, decision tree, Bayesian classifier, etc.

Hachesu et al. (2013) applied techniques such as SVM, decision tree and ANN to predict the duration of stay of patients. J. Kim et al. (2015) used a Classification and Regression Tree [ART] (an algorithm of the fuzzy logic and decision tree) to design a CHD detection model using data sets derived from the Korean National Health and Nutrition Examination Survey VI (KNHANES-VI). H. Kim, Ishag, Piao, Kwon and Ryu (2016) proposed a prediction system for CVD using ANN and the Bayesian classifier based on Heart Rate Variability (HRV) and carotid images. These systems (Hachesu et al., 2013; J. Kim et al., 2015; Kumari & Godara, 2011; Melillo et al., 2015) based on data mining techniques were developed for classification and clustering purpose. They can give a classification result if an individual would develop a CVD event or not but an absolute risk score of potential CVD disease cannot be estimated.

Some CVD models can either provide a predicted risk level (Vaanathi, 2017) or a definite risk score (Unnikrishnan et al., 2016; Murukesan et al., 2014) to patients. A CVD diagnosis model based on neuro-fuzzy expert system (Vaanathi, 2017) combined the neural networks and adaptive neuro-fuzzy for CVD risk estimation. After demanded fields are inputted, the risk level (very low, low, high) will be given. Murukesan et al. (2014) used Probabilistic Neural Network (PNN) to predict the risk of Sudden Cardiac Arrest (SCA). The risk model proposed by Unnikrishnan et al. (2016) has the ability to compute risk scores for patients and be a supportive tool for diagnosis of CVD and the decisions of cardiologists. However, only a definite risk score can be estimated, the duration of time for the estimated risk is unknown.

In summary, modelling methods used to analyse the risk in CVD prediction are traditional regression analysis, data mining techniques (classification and clustering, and Artificial Intelligence (AI)). Table 2.2 summarises the modelling methods reviewed as well the merits and drawbacks for a specific method.

Table 2.2: Modelling Methods in CVD Risk Prediction

Methods	Sub-categories	Individual Methods	Merits	Drawbacks
Statistical Regression Analysis	Parametric Approaches	Exponential	Be useful when the samples are small	Not enough to use standard methods to estimate the results. Need to fulfill a proportional hazard assumption.
		Weibull		
		Gamma		
	Semi-parametric Approaches	Cox proportional hazards regression analysis	Works well for both quantitative predictor variables and categorical variables. Investigating simultaneously the importance of various risk factors on survival time	Need to fulfill an assumption of a linear association between the natural logarithm of the relative hazard and the predictors
	Non-parametric Approaches	Kaplan-Merier	Works well when the variable is categorical	Only shows the effect under one factor, not useful when the input variable is quantitative
		Log-rank test		
Data Mining	Classification & Clustering	Support Vector Machine (SVM)	Higher sensitivity and specificity compared with statistical methods	Classification or clustering results are obtained, or risk of different levels is given, no accurate risk score
		Decision tree		
		Bayesian classifier		
		RIPPER classifier		
	Probability Estimation	Probabilistic Neural Network (PNN)		An ultimate risk can be estimated. Cannot know how many years for the estimated risk
		Artificial neural networks (ANNs)		
		Fuzzy Logic		

2.3.3 CVD Risk Factors

2.3.3.1 Overview of CVD Risk Factors

Apart from the modelling methods, the inclusion of risk factors also decides the performance of a prediction model. A review paper conducted by Damen et al. (2016) suggests that the number of predictors that have been included in CVD models is greater than 100. The top 20 are listed in Table 2.3.

Conventional risk predictors include sex, smoking, age, SBP, total cholesterol, diabetes, hypertension, HDL cholesterol and Body Mass Index (BMI), etc. (Cupples, 1987). According to Table 2.3, most items in the list are traditional factors, which indicates that the majority of studies on CVD prediction relied on traditional predictors for risk estimation.

Some researchers have tried to explore novel risk factors such as kidney function, heart rate, family history, C-reactive Protein (CRP). Odden et al. (2014) combined traditional risk factors (including SBP, HDL cholesterol, Low Density Lipoprotein (LDL) cholesterol, obesity, and diabetes) and novel risk factors (including N-terminal

Table 2.3: Top 20 Predictors in CVD Models

(Damen et al., 2016)

Ranks	Factors
1	Smoking
2	Age
3	Systolic blood pressure
4	Total cholesterol
5	Diabetes
6	Hypertension
7	HDL cholesterol
8	Body mass index
9	Sex
10	Family history of CVD
11	Electrocardiography
12	Race
13	Total: HDL cholexterol ratio
14	Heart Rate
15	Blood glucose
16	Physical condition
17	Diastolic blood pressure
18	Previous CVD
19	Non-HDL cholesterol
20	LDL cholesterol
Notes: HDL=high density lipoprotein; LDL=low density lipoprotein.	

pro-B-type natriuretic peptide, C-reactive protein, and kidney function) to predict CVD risk. A neuro-fuzzy system recently designed by Vaanathi (2017) included four novel risk factors (heart rate, chest pain type, blood sugar, and exercise) and four traditional predictors (sex, age, cholesterol, and blood pressure).

2.3.3.2 Number of Risk Factors in CVD Model

CVD models can be developed based on a single risk predictor or multiple risk predictors. The majority of studies (Lloyd-Jones et al., 2004; Assmann et al., 2002; Conroy et al., 2003; Assmann et al., 2002; Hippisley-Cox et al., 2007; Vaanathi, 2017; Y.-M. Liu, Chen, Yen & Chen, 2013) developed models based on multivariate analysis. In multiple-predictor models, seven or eight is the median number of factors included.

D'Agostino et al. (2008) developed two sex-specific risk algorithms for the prediction of CVDs. One is based on traditional risk factors including age, anti-hypertensive medication treatment for hypertension, total cholesterol, HDL cholesterol, current status of smoking, SBP, and diabetes. The other one is based on non-laboratory test risk factors and replaced the total cholesterol and HDL cholesterol with BMI. A health parameter model using SVM for CVD prediction by Unnikrishnan et al. (2016) included same predictors as the FHS laboratory-test-based model (D'Agostino et al., 2008). The QRISK study researchers included eight risk factors in their model, the first seven of them were the same as the FHS model and the last one is family history (Hippisley-Cox et al., 2007). A neuro-fuzzy system designed by Vaanathi (2017) has eight input parameters including sex, age, cholesterol, blood pressure, heart rate, chest pain type, blood sugar, and exercise.

Researchers have been working on exploring novel predictors and extending the number of predictors for estimation of CVD risk. The cohort study conducted by De Ruijter et al. (2009) incorporated predictors included as the Framingham risk equation and four new bio-markers (interleukin 6, CRP, folic acid, and homocysteine) to detect CVD risk in the elderly people. The risk factors has been increased to 11. This model obtained an improvement of accuracy for identification of high risk of CVD over the FHS model. A Bayesian clinical reasoning model for CVD prediction (Y.-M. Liu et al., 2013) was developed by incorporating 12 risk factors including demographic

features (gender, age and family history), metabolic syndrome components (betel quid (male only), fasting glucose, HDL-cholesterol, triglyceride, waist circumference and metabolic score) and conventional predictors (alcohol, SBP and diastolic blood pressure).

The multiple variables CVD prediction models that have been reviewed are summarised in Table 2.4

Table 2.4: Summary of CVD Prediction Models Using Multiple Risk Factors

CVD Prediction Models Based on Multiple Varaiables	Num. of Predictor	Predictors in the Model
The FHS laboratory-test-based model by D'Agostino et al. (2008)	7	Age, total cholesterol, HDL cholesterol, SBP, anti-hypertensive medication treatment for hypertension, current smoking and diabetes status
The FHS non-laboratory-based model by D'Agostino et al. (2008)	6	Age, BMI, SBP, anti-hypertensive medication treatment for hypertension, current smoking and diabetes status
The SVM health parameter model by Unnikrishnan et al. (2016)	7	Age, total cholesterol, HDL cholesterol, SBP, anti-hypertensive medication treatment for hypertension, current smoking and diabetes status
The QRISK model by Hippisley-Cox et al. (2007)	8	Age, total cholesterol, HDL cholesterol, SBP, anti-hypertensive medication treatment for hypertension, current smoking, diabetes status, family history
The neuro-fuzzy-based model by Vaanathi (2017)	8	Sex, age, cholesterol, blood pressure, heart rate, chest pain type, blood sugar, and exercise
The CVD prediction model by De Ruijter et al. (2009)	11	Age, total cholesterol, HDL cholesterol, SBP, anti-hypertensive medication treatment for hypertension, current smoking, diabetes status, homocysteine, folic acid, C reactive protein, and interleukin 6
The Bayesian clinical reasoning model by Y.-M. Liuet al. (2013)	12	Gender, age, family history, betel quid (male only), fasting glucose, HDL-cholesterol, triglyceride, waist circumference, metabolic score, alcohol, SBP and diastolic blood pressure

Apart from models based on multiple risk factors, there are some models targeting the investigation of CVD risk prediction based on the analysis of a single risk parameter. The Kailuan study researchers (Yu et al., 2017) generated a risk prediction model based on the cumulative exposure to Resting Heart Rate (cumRHR). Murukesan et al. (2014) employed a machine learning approach to detect Sudden Cardiac Arrest (SCA) by analysing HRV. In a study for estimating the risk of 10-year Atherosclerotic Cardiovascular Disease (ASCVD), the same risk factor HRV in the SCA detection

model (Murukesan et al., 2014) was inputted to the model (Han et al., 2017).

2.4 Data Availability and Utilisation for CVD Research

2.4.1 Data Features and Study Population

A great number of CVD data sets have been created by investigators all over the world. It is reviewed that data sets used for the development of CVD prediction models are mostly from a longitudinal cohort study (Damen et al., 2016). A longitudinal cohort study is a long-term population research. Normally, the length of data collections of these studies will be across many years.

Starting times of longitudinal cohort studies that have been done are quite different. The famous FHS in the United States began than the others, the data collection starting in 1948. The collected data were firstly used for the identification of risk predictors for heart disease (Kannel, Dawber, Kagan, Revotskie & Stokes, 1961). The Framingham researchers still kept focusing on heart-related research (Kannel et al., 1979; Splansky et al., 2007). The earliest cohort study on CVD detection in Asia was started in Japan (Collaboration et al., 2004). In 1961, the Japan Hisayama cohort study (Hasuo et al., 1989) was started with the diagnosis on death certificates referring to CVD in Hisayama, Japan.

The study population has diversity as well. Figure 2.2 summarises the distribution of cohort study populations in different continents (Damen et al., 2016). According to this figure, approximately half of the CVD longitudinal cohort studies occurred in Europe (46%). 36% of the study population were from the United States and Canada. Those from Asia and Australia were 12% and 4% respectively. The remaining 2% is cross-continental. Table 2.5 lists the well-known studies from different study populations.

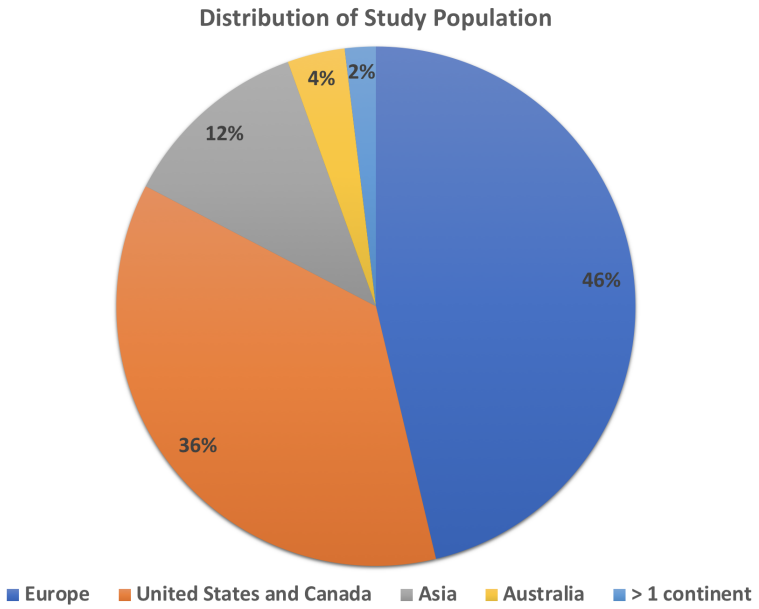


Figure 2.2: Distribution of the Population in Longitudinal Cohort Studies (Damen et al., 2016)

Table 2.5: Well-known Studies from Different Populations

Continents	Studies
Europe	The SCORE Cohort Study (Conroy et al., 2003)
United States	The Framingham Cohort Study (Lloyd-Jones et al., 2004)
Canada	The Study of HealthAssessment and Risk in Ethnic groups (SHARE) (Anand et al., 2000)
Asia	The Seven Cities cohort study in China (M. Liu et al.,2007 The KMIC in South Korea (Jee, Suh, Kim & Appel, 1999) The Japan Hisayama Cohort Study (Hasuo et al., 1989)
Australia	The Melbourne Cohort Study (Harriss et al., 2007)

2.4.2 Content of Data Sets

Apart from the study populations and the starting times, there are noticeable differences in the data contents between different CVD longitudinal cohort studies. The content of a data set refers to two aspects: risk factors and sample size (Damen et al., 2016).

Differences of risk factors among different cohort studies might be caused by the characteristics of populations including differences of age, sex, race, etc. For example,

the data of the FHS (Mahmood, Levy, Vasan & Wang, 2014) was collected from the population in the United States but the Keelung Community-based Integrated Screening (KCIS) study (Chen et al., n.d.) was started in Asia, so the data features regarding risk factors are different because of the differences of the population features. Apart from this, the data of clinical markers might be different as well because each study has its individual research objectives and corresponding design.

The number of participants involved in a cohort study data collection is called "sample size" and that needs to be considered in a data analysis. A systematic review in 2016 indicates that the sample size in CVD cohort studies ranges from 51 to 1189845 and the median number is 3969 (Damen et al., 2016). Around 14,000 people attended the FHS. The number of attendants in the KCIS study was 61,869 (Chen et al., n.d.).

These two aspects should be considered when researchers look for research data sets. Researchers need to firstly figure out the risk factors included in a data set and choose the right one. A larger sample size is beneficial for mining effective data patterns and useful information (F. Harrell, 2013).

2.4.3 The FHS Data Set

This data set originated from the FHS in 1948. It is a world-class, long-time, ongoing cardiovascular cohort study directed by the National Heart, Lung and Blood Institute (NHLBI) (Grundy et al., 2005). 5209 adult subjects aged between 30 and 62 who had not been diagnosed as CVD with related overt symptoms previously, were recruited at the beginning of this study (Mahmood et al., 2014).

The origin of the Framingham Research is related to the death of President Franklin D. Roosevelt who suffered from hypertensive heart failure but without being diagnosed and treated timely (Mahmood et al., 2014). Thus, the FHS scientists started this project to investigate the risk factors that contribute to heart disease. Today, the Framingham

study has become the epicenter for CVD, bone, and sleep research (Mendis, 2010).

Over 60 years, the FHS has gone through several "cohorts", the Original Cohort in 1948, the Offspring Cohort in 1971, the Omni Cohort in 1994, the Third Generation Cohort in 2002, the New Offspring Spouse Cohort in 2003, and the Second Generation Omni Cohort in 2003. A summary of each cohort is listed in Table 2.6.

Table 2.6: Summary of The FHS Data Set

Cohorts Names	Years	Numbers of Participants	Ages	Gender	Participants Recruited
The Original Cohort	1948	5,209	30 - 62	Male & Female	Randomly sample from 2/3 of the adult population of Framingham, Massachusetts, without history of heart attack or stroke
The Offspring Cohort	1971	5,124	Unknown - 70	Male & Female	The offspring of the Original Cohort and their spouses
The Omni Cohort	1994	507	20 - 79	Male & Female	New group of participants of African-American, Hispanic, Asian, Indian, Pacific Islander and Native American origins, who were residents of Framingham and the surrounding towns
The Generation Three Cohort	2002	4,095	19 - 79	Male & Female	Third generation of participants from Offspring Cohort
The New Offspring Spouse Cohort	2003	103	40 - 89	Male & Female	Spouse of an Offspring who were never enrolled in the Framingham Heart Study
The Second Generation Omni Cohort	2003	410	20-70	Male & Female	New participants related to the participants of Omni Cohort 1 and also individuals unrelated to Omni Cohort 1 members

The FHS data set has provided valuable data resources for researchers to investigate the identification and prevention of CVD all over the world (Mendis, 2010). It was started with identifying key risk predictors of CHD such as obesity, diabetes, levels of blood cholesterol, smoking, exercise habit, blood pressure, etc. (Kannel et al., 1961). Based on that, researchers started to explore CVD risk factors, as well as how these factors had affected the development of CVD since the 1960s (Kannel, Abbott, Savage & McNamara, 1982). The newly generated Third Generation Cohort and New Offspring Spouse Cohort will keep providing contributions to the research on CVD (Parikh et al., 2007).

2.4.4 The KCIS Programme Data Set

This data set was generated along with the KCIS programme (Chen et al., n.d.). It is a screening programme targeting multiple diseases. Five types of neoplastic disease (colorectal cancer, breast cancer, oral cancer, cervical cancer, and liver cancer) and three types of chronic conditions (type 2 diabetes, hyperlipidemia, and hypertension) were screened. One objective of this study was to create a health information system for screening cancers and chronic diseases.

The KCIS programme was started from 1999 and ended in 2003. The Health Bureau of Keelung City was responsible for the execution of this programme. In the beginning, the target population was 217,895 residents living in Keelung. In total, 61,869 attendants aged 30 to 79 years from different areas (Joshang, Jongjeng, Shinnyi, Renay, Noannoun, Anleh and Chiduu) in Keelung participated in this programme until the end of 2003 (Chiu et al., 2006). A summary of the participants in the KCIS study is listed in Table 2.7.

Table 2.7: Summary of the KCIS Data Set

(Chiu et al., 2006)

Age Group	Target Population			Attendants			Coverage
	Female	Male	Total	Female	Male	Total	Rate
30 - 39	34468	36255	70723	8935	4334	13269	18.8
40 - 49	32371	33371	65742	10855	6067	16922	25.7
50 - 59	18016	17440	35456	7944	4395	12339	34.8
60 - 69	13646	12554	26200	7633	4421	11154	42.6
70 - 79	8697	11077	19774	4040	4145	8185	41.4
Total	107198	110697	217895	38507	23362	61869	28.4

2.5 eHealth Solutions for CVD Prediction

eHealth mainly refers to the digital health which uses information technology to provide convenient health care management. It was first defined by the International Telecommunication Union (Eng, 2001). With the rapid development of telecommunications

and computer technologies, a lot of eHealth solutions have been developed for CVD prediction, including clinical guidelines and standards (Cardiovascular Disease Risk Assessment Steering Group and others, 2017; CG181, NICE, 2014; Expert Panel on Detection et al., 2001), online tools (Framingham Heart Study, 2017; Wells et al., 2017; Wells, Kerr, Eadie, Wiltshire & Jackson, 2010), and mobile health care tools (Patrick et al., 2009; Carter, Burley, Nykjaer & Cade, 2013). It is expected that these solutions can not only provide suitable health care but also help people with optimal cardiovascular health management.

2.5.1 Clinical Guidelines

Clinical guidelines are important tools for health care practitioners for cardiovascular risk assessment and cardiovascular risk factor management. A CVD guideline can be introduced on the basis of a systematic evidence review or promoted prediction models or a combination of both. Heath care practitioners can effectively use these guidelines to identify potential CVD patients and give corresponding treatments.

A guideline by the National Institute for Health and Clinical Excellence (NICE) in July 2014 was developed based on a systematic evidence review (CG181, NICE, 2014). It covers the risk assessment of CVD for potential patients and provides care guidance for adults who have CVD. Rules for the identification of CVD risk, recommendations on the lifestyle intervention and treatments for lipid modification are listed. The Third Report of the Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) (Expert Panel on Detection et al., 2001) is a CVD guideline book incorporating the Framingham CHD prediction algorithms (Wilson et al., 1998). The New Zealand Primary Care Handbook (Cardiovascular Disease Risk Assessment Steering Group and others, 2017) is a guideline promoted for New Zealanders based on evidence reviews and promoted algorithms. The cardiovascular

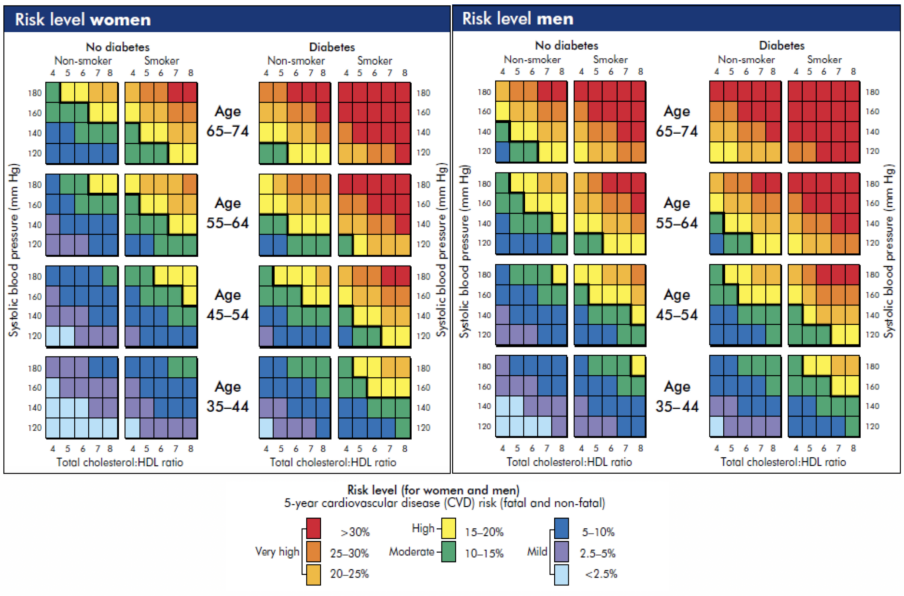


Figure 2.3: New Zealand Cardiovascular Risk Charts for 5-year Risk of CVD Prediction (Cardiovascular Disease Risk Assessment Steering Group and others, 2017)

risk charts for risk prediction were introduced referring to the old Framingham CVD prediction algorithms (Wilson et al., 1998). To lower the risk of developing CVD in five years, general, specific, or intensive lifestyle interventions were introduced. In addition, other treatments such as smoking interventions, complementary therapies, lipid modification, and blood pressure lowering were also listed thoroughly. Figure 2.3 is a graph of New Zealand Cardiovascular risk charts for the 5-year risk of CVD prediction.

2.5.2 Web-based Tools

FHS investigators have done a lot of research on cardiovascular and heart-related disease. Risk prediction calculators for specific CVDs and the general CVD were developed based on the achievement of these studies (D'Agostino et al., 2008). They can be applied to congestive heart failure, CHD, diabetes, intermittent claudication, stroke, and general CVD. These calculators are easy to operate. After values of risk factors are inputted a specific risk score (probability of occurrence of the disease) for an individual

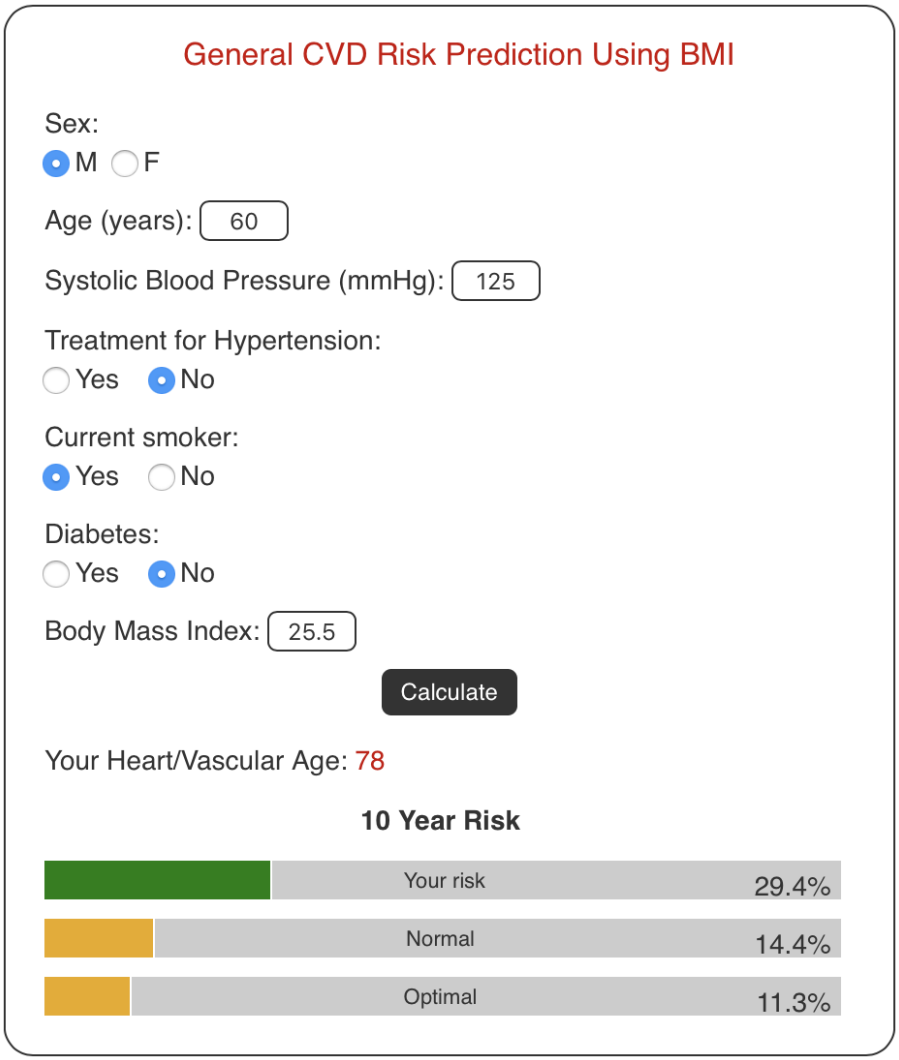


Figure 2.4: The Framingham 10-year General CVD Risk Prediction Tool (D'Agostino et al., 2008)

will be given. Additionally, they can be conveniently accessed online, so they have become very popular risk estimation tools among physicians and individuals. Figure 2.4 is a screen-shot of the Framingham web-based risk prediction tool for general CVD using non-laboratory predictors.

Cardiovascular Risk Charts and Calculator is a web-based tool developed by the University of Edinburgh (Wells et al., 2017). Cardiovascular risk calculators provided for health care professionals are based on formulas from the FHS (D'Agostino et

al., 2008), the Joint British Societies (JBS) study (Board, 2014), the British National Formulary (BNF) (Mehta, 2005), and the ASSIGN study (Woodward et al., 2006). Similarly, the risk of developing CVD, CHD, stroke, etc. will be calculated and displayed in different graphic styles, BNF Charts, smiley faces, comparison bars, or thermometers.

Your Heart Forecast is an on-line platform jointly developed by the Heart Foundation, University of Auckland and Enigma (Wells et al., 2010). The algorithm was derived from the New Zealand Primary Care Handbook (Cardiovascular Disease Risk Assessment Steering Group and others, 2017). This tool has two versions. One is for New Zealanders only. The other one is for international use. Heart age and stroke risk in the next five years will be calculated after answering related questions. Other available online tools for CVD risk assessment are PREDICT (Wells et al., 2017), Best Practice (Gill & Mangin, 2011), and so on.

2.5.3 mHealth Applications

mHealth application is a subset of eHealth solutions for decision makers in the assessment of CVD. It was originally promoted by the Global Observatory for eHealth (GOe) of the World Health Organisation (WHO) and was defined as the "medical and public health practice supported by mobile devices, such as mobile phones, patient monitoring devices, Personal Digital Assistants (PDAs), and other wireless devices" (World Health Organisation and others, n.d.). In recent decades, health care applications based on mobile phone and wearable sensors have become popular.

2.5.3.1 Mobile Phone Health Care Applications for CVD

With the rapid development of mobile phone technologies, a great number of mobile-based applications have emerged for the prevention and prediction of CVD. Mainly,

two types of mobile-based applications are common in the literature and the consumer app stores. The major mobile applications focus on lifestyle intervention and self-management for the prevention of CVD. The others are decision support systems for the risk prediction of CVD.

Obesity is a factor directly leading to CVD (Wilson, D'agostino, Sullivan, Parise & Kannel, 2002). Taking the weight management interventions as an example, a message-based interventions programme relying upon mobile communication techniques was proposed for weight loss (Patrick et al., 2009). Advice on lifestyle behaviours was sent to the overweight groups via mobile messages. Another smart-phone application, called My Meal Mate (MMM), can provide self-monitoring management for daily activities (Carter et al., 2013). Users can set their goals of diet and activities in this app and then feedback of the daily lifestyle changes will be recorded. Similar applications are used to increase physical activities (Hurling et al., 2007), to improve hypertension care (Green et al., 2008), to quit smoking (Whittaker et al., 2012), and to control elevated blood pressure (Burke et al., 2015). Compared with web-based tools and paper diaries, smartphone-based apps are much more convenient for lifestyle interventions, but they cannot accomplish the risk prediction of CVD.

Compared to applications introduced above, the Stroke RiskometerTM app (Parmar et al., 2015) is an effective tool for the risk estimation of stroke. Pre-designed questions regarding stroke risk predictors are firstly asked sequentially, and then a risk score is given. One drawback of the Stroke RiskometerTM is the lack of real-time monitoring of risk factors. Another mobile application incorporating the SCORE algorithms (Conroy et al., 2003) can calculate CVD risk from monitoring of blood pressure in combination with other clinical factors. Measures of blood pressure are obtained by a pressure monitor and then are transmitted to an Android application installed in a smartphone via Bluetooth. The system presented by Hervás et al. (2013) has achieved the real-time monitoring risk of developing CVD for an individual but only a risk level can be

obtained from this system, as shown in Table 2.8.

Table 2.8: Risk Levels in Real-time Monitoring System

(Hervás et al., 2013)

Risk Levels	Range of Risk
Very High	If a user presents a risk of 15% and over
High	If the risk is in the range 10%–14%
Mid High	User presents a risk from 5% to 9%
Mid	User presents a risk from 3% to 4%
Mid Low	If the risk is 2%
Low	If the risk presented corresponds to 1%
None	No risk is presented

2.5.3.2 Wearable Sensor Monitoring System for CVD

Wearable sensors are mobile devices that can be worn for a period of time (Pantelopoulos & Bourbakis, 2010). They can monitor many dimensions of real-time data such as exercise, electrocardiography (ECG), heart rate, breathing rate, pulse pressure, temperature etc. These real-time data can be reviewed and processed remotely later for health care or decision making. Patel, Park, Bonato, Chan and Rodgers (2012) developed a remote monitoring system based on a wearable sensor for primary health care. In this system, motion and physiological information are collected via a body-worn sensor, and then is processed and synchronised to caregivers via a communication gateway such as a cell phone. Caregivers can implement health care interventions according to the information transferred.

With the recent progress of technologies in body-worn devices, smart textiles, wireless communications and micro-electronics, the development of monitoring systems (Jin, Oresko, Huang & Cheng, 2009; Oresko et al., 2010; Lin, Yang, Wang & Yang, 2012) based on wearable sensors has made great advances for health care and clinical

decision making. The continuous advance of these systems will possibly revolutionise the future of health care by encouraging the involvement of both physicians and individuals with personal and ubiquitous monitoring devices.

Research on the wearable monitoring system for CVD mostly focuses on sensors and data collection (Etemadi et al., 2016; Milenković, Otto & Jovanov, 2006), connectivity between devices and visualisation platforms (W. Lee, Yoon & Park, 2016; Michard, 2017), textiles (Centers for Disease Control and Prevention and others, 2009; Choi & Jiang, 2006), and applications for self-management and treatment (Milani & Franklin, 2017). The medical practice of health care monitoring systems based on wearable sensors for CVD risk predictions is still under-utilised due to the limited capacity of data processing. Barriers encountered are the low efficiency of data processing such as algorithms with unsatisfactory accuracy for clinical decision making or increased data processing time.

Some researchers have tried to process and utilise the data collected for CVD detection. Wearable sensor-based platforms developed by Oresko et al. (2010) can provide diagnostic solutions for CVD estimation. Electrocardiography (ECG) real-time data acquired via portable Holter monitors is firstly extracted and then processed using classification algorithms. Another wearable sensor monitoring system used for CVD health care (Lin et al., 2012) integrated two neural network classification algorithms (radial basis function network (RBFN) and generalised regression neural network (GRNN)) into the estimation of everyday energy expenditure. Systems reviewed above can give a risk level to develop CVD and show an assistance for CVD detection. However, classification algorithms used in the prediction models cannot provide accurate scores. Data processing algorithms that can give accurate risk estimation of CVD are barely used. Widespread integration of this technology into CVD detection continues to be limited. Articles reviewed on eHealth applications are summarised in Table 2.9.

Table 2.9: Lists of eHealth Solutions

eHealth Solutions	Literature Reviewed
Clinical Guidelines	The New Zealand Primary Care Handbook (Cardiovascular Disease Risk AssessmentSteering Group and others, 2017).
	Cardiovascular disease: risk assessment and reduction (CG181 & NICE, 2014)
	The Third Report of the Expert Panel on Detection, Evaluation, and Treatment of HighBlood Cholesterol in Adults (Expert Panel on Detection, Evaluation and others, 2001).
Web-based Tools	Risk Score Calculators from Framingham Heart Study (D'Agostino et al., 2008).
	Cardiovascular Risk Charts and Calculator Based on FRS BNF ASSIGN (The University of Edinburgh, 2017).
	Your Heart Forecast (Wells et al., 2010).
Mobile Phone Health Care for CVD	Texting-based Interventions Program (Patrick et al., 2009).
	My Meal Mate (MMM) (Carter et al., 2013).
	Stroke RiskometerTM App (Parmar et al., 2015).
Wearable Sensor Monitoring System for CVD	Heart-to-go (Jin, Oresko, Huang & Cheng, 2009)
	Wearable Smartphone-based Platform for Real-Time CVD Detection ms (Oresko et al., 2010).
	Wearable Sensor Module for Daily Energy Expenditure Estimation (Yang,Wang & Yang, 2012).

2.6 Summary

In summary, this section reviewed the previous work towards to CVD prediction, CVD prediction models, data availability and utilisation, and CVD eHealth applications. By having an overview of CVD predictions, we have a general understanding about the CVD outcomes that previous works targeted. Most of the research focused on specific CVD components like stroke and heart attack. Guided by the CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies (CHARMS) developed by Moons et al. (2014), we also reviewed CVD studies in terms of CVD prediction models and modelling methods. After that, we researched risk factors and CVD research data sets related to CVD risk predictions. We had a deep review into two available CVD research data sets: the Framingham Heart Study data set

(Mahmood et al., 2014) and the KCIS data set (Chiu et al., 2006). Lastly, we reviewed guidelines and applications for CVD predictions.

Chapter 3

Research Methodology

3.1 Introduction

The design of research method is to choose suitable research approaches to conduct a research. Defining the plans and procedures of doing a research previously helps the researchers to obtain the detailed methods of collecting data, doing data analysis, and interpreting results of analysis from primitive assumptions (Leavy, 2017). It is critical to a rigorous research.

To decide what research approaches were suitable for our research, we needed to consider three components involved in a research: the philosophical worldview, research approaches and specific research methods (Creswell, 2014).

In this chapter, we will first have a look at the research design, including philosophical worldviews and research methods. Then, we will select a research method that is consistent with the expectations of our research and its fundamental characteristics. After determining the appropriate research method, we will introduce the procedures for implementing the research method. Specific steps regarding the data collection and data analysis will be stated.

3.2 Research Design

Research designs are defined as "strategies of inquiry" (Denzin & Lincoln, 2011). In other words, they are specific directions to do a research. Each research approach has its own research designs which lead the researcher to select a specific research method. Philosophical worldview decides the latter so we will start with it.

3.2.1 Philosophical Worldview

Philosophical worldview (Guba, 1990), also called "broadly conceived research methodologies" (Neuman, 2002), is defined as "a basic set of beliefs that guide actions". It is regarded as a hidden factor that influences the practice of a research and helps us find out which research approaches should be chosen (Creswell, 2014). Four worldview positions are promoted by Creswell (2014): postpositivism, constructivism, transformative, and pragmatism. Each worldview has its own elements and characteristics.

Postpositivism worldview is the traditional form of doing research (Smith, 1983). Major elements are determination, reductionism, empirical observation and measurement, and theory verification (Creswell, 2014). A research holding a postpositivism worldview is more likely to be quantitative research than qualitative research. This kind of research typically begins with a theory, collecting data, and then doing data analysis to approve or refute the theory (Phillips & Burbules, 2000). If necessary, revision and an additional test will also be conducted.

Constructivism worldview typically leads to be a qualitative approach held by social constructivists. Rather than testing a theory as in postpositivism, it intends to interpret the meanings that individuals put on the world (Mertens, 2014).

Transformative worldview arose because some researchers felt that postpositivism and constructivism worldviews could not fulfil the needs of marginalised people or address issues of politics such as power, social justice, and discrimination (Fay, 1987).

The philosophical worldview cares about the needs of people in our society that may have been treated unequally or disenfranchised (Mertens, 2014).

Pragmatic worldview doesn't focus on one specific research approach but emphasises the research questions and tries all potential research methods to understand the research questions (Rorty, 1990). Thus, it is a typical worldview of mixed methods studies.

According to definitions of the four philosophical worldviews, the postpositivism worldview fits this research. A characteristic of our research is figuring out the theory of cause and effect on the development of CVD. Guided by this worldview, a proper research approach will be selected.

3.2.2 Research Approach

Creswell (2014) classified research approaches into three categories, i.e. qualitative, quantitative, and mixed methods as described in Figure 3.1.

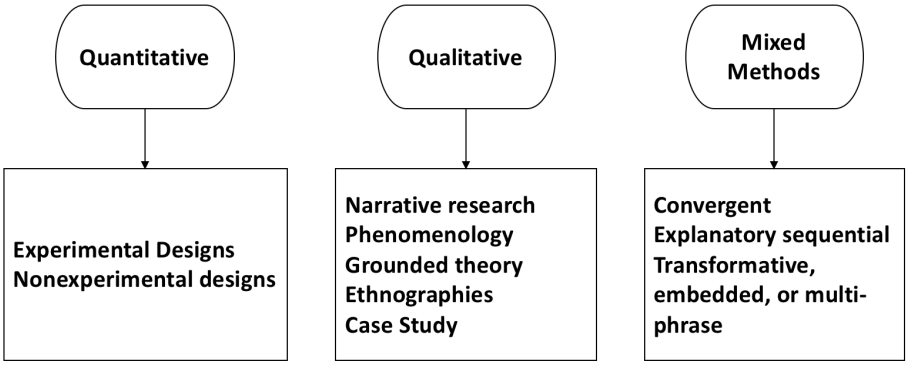


Figure 3.1: Overview of Research Design within Research Approaches (Creswell, 2014)

Quantitative research is an approach to systematically investigate objective theories via statistical or mathematical techniques, and then to test the theories and hypotheses

pertaining to the phenomena (Given, 2008). It is an empirical exploration of observable phenomena by examining the relationships between variables deductively (Creswell, 2014). The objective of this approach is to determine causes and effects or to describe some attributes among a population. The data of quantitative research is normally in numeric form because this research mainly focuses on "how much" or "how many" (Merriam & Tisdell, 2016).

Qualitative research is the opposite of quantitative research. It is an approach to uncover and understand the meanings of a phenomenon or situation or event for some individuals (groups) involved (Creswell, 2014). Qualitative research uses words as data form. Data can be collected in many ways. For example, when collecting data using interviews, potential methods could be unstructured, semi-structured, structured, observations, reflective notes, focus groups, photographs, videos, texts, and so on (Savin-Baden & Major, 2013).

Mixed methods research, also called integrated methods research, is an approach that collects both quantitative and qualitative data. It integrates these two types of data for the data analysis by employing either quantitative approaches, or qualitative approaches, or both in a single study (Tashakkori & Teddlie, 2010). It is commonly used in some complex situations where researchers cannot answer research questions or assumptions using a single method. Or they may want to investigate different aspects of the same phenomenon. Thus, more than one research approach needs to be demanded, both quantitative and qualitative (Morse & Niehaus, 2009), and data regarding these two approaches will be collected (Morse & Niehaus, 2009).

In this research, the theory we are trying to determine is whether early detection of CVD could help physicians and individuals improve the health condition effectively using a wearable device, and further reduce the probability of the occurrence of CVD in the next certain years. The causes will be predictors that might contribute to the occurrence of CVD. Data will be in numerical form, such as the age, sex, heart rate,

and so on.

We want to research the relationship between risk factors and the development of CVD deductively. By matching philosophical worldviews and research approaches introduced above to the characteristic of this research, quantitative approach is suitable for this study.

The final step of doing a research within a research framework is to define a process for executing the research approach chosen previously, including data collection, data analysis, results interpretation and validation. It is useful to consider all these procedures thoroughly before conducting a research. Identified as a quantitative approach research, we will focus on the design of quantitative research.

3.3 Data Collection

3.3.1 Choice of Data Set

As reviewed in Chapter 2, many CVD related data sets are available (M. Liu et al., 2007; Jee, Suh, Kim & Appel, 1999; Harriss et al., 2007; Lloyd-Jones et al., 2004). Regarding three aspects reviewed in Section 2.4, the feature of a data set, the starting time, and the data set content, the FHS data set (Mahmood et al., 2014; Dawber, Kannel & Lyell, 1963; Kannel et al., 1979) was selected for developing the prediction model.

The first reason we chose the FHS data set is that it is a longitudinal and etiological study, which is extremely suitable for survival analysis. A long term prospective data collection was conducted. CVD prediction can be categorised as one type of survival analysis. This feature of the FHS data set was extremely conducive for us doing this research.

Considering the starting time of a data set, the FHS started in 1948, so more than 60 years follow-up data were collected. The FHS is the first prospective study on CVD and

has become a landmark in the epidemiology field. Over the past decades, approximately 3000 articles using this data set have been published in leading medical journals.

Last but not least, the contents and the sample size of the FHS data set is sufficient for us to develop an effective prediction model. Data regarding demographics, family history, lifestyle, disease history, physical examination, lab test, etc. have been gathered and recorded in numeric form. In addition, over 14,000 people covering three generations (the original participants, the offspring of the original participants, as well as their grandchildren) have participated this study. To fit a reliable prediction model, the desirable sample size must be larger than 10 or 20 times the number of candidate risk factors (J. F. Harrell, Lee, Matchar & Reichert, 1985). This hypothesis is far more supported.

The Original Cohort study data set is enclosed for this research as it has the largest number of samples. This data set comprises 5209 subjects aged 30 to 62 years who had not symptoms of CVD as the baseline. The distributions of age and sex are summarised in Table 3.1. Data were collected at the town of Framingham in Massachusetts, by which the data set was named. Totally 32 exams data were collected, as listed in Table 3.2.

Table 3.1: Age and Sex Distributions of the Framingham Original Cohort Study

Age	Men	Women	Totals
29-39	835	1,042	1,877
40-49	779	962	1,741
50-62	722	869	1,591
Totals	2,336	2,873	5,209

However, data from 130 subjects were removed from the data set we requested because they didn't agree to publishing of their data. Finally, we had 5079 observations for our model fitting, and 3189 events (CVD) occurred.

Table 3.2: Exams in the Framingham Original Cohort Study

Exams	Exam Date Range	Age Range	Mean Age	Attendees
Exam 1	1948 - 1953	28 - 74	44	5209
Exam 2	1950 - 1955	31 - 65	46	4792
Exam 3	1952 - 1956	32 - 67	48	4416
Exam 4	1954 - 1958	34 - 69	50	4541
Exam 5	1956 - 1960	37 - 70	52	4421
Exam 6	1958 - 1963	38 - 72	54	4259
Exam 7	1960 - 1964	40 - 74	55	4191
Exam 8	1962 - 1966	42 - 76	57	4030
Exam 9	1964 - 1968	44 - 78	59	3833
Exam 10	1966 - 1970	46 - 80	61	3595
Exam 11	1968 - 1971	49 - 81	62	2955
Exam 12	1971 - 1974	50 - 83	64	3261
Exam 13	1972 - 1976	53 - 85	66	3133
Exam 14	1975 - 1978	55 - 88	68	2871
Exam 15	1977 - 1979	57 - 89	69	2632
Exam 16	1979 - 1982	59 - 91	70	2351
Exam 17	1981 - 1984	61 - 93	72	2179
Exam 18	1983 - 1985	63 - 94	74	1825
Exam 19	1985 - 1988	65 - 96	75	1541
Exam 20	1986 - 1990	67 - 97	77	1401
Exam 21	1988 - 1992	69 - 99	79	1319
Exam 22	1990 - 1994	72 - 101	80	1166
Exam 23	1992 - 1996	73 - 101	81	1026
Exam 24	1995 - 1998	76 - 103	83	831
Exam 25	1997 - 1999	78 - 104	84	703
Exam 26	1999 - 2001	79 - 103	86	558
Exam 27	2002 - 2003	82 - 104	87	414
Exam 28	2004 - 2005	84 - 104	89	303
Exam 29	2006 - 2007	85 - 102	91	218
Exam 30	2008 - 2010	88 - 102	92	141
Exam 31	2010 - 2011	90 - 99	92	91
Exam 32	2012 - 2014	93 - 106	96	40

3.3.2 Process of Data Collection

A subset of the FHS data set used for teaching and research purposes is available on the website of NHLBI (National Heart, Lung, and Blood Institute, 2018). We requested it with the approval of the Auckland University of Technology Ethics Committees and an agreement with NHLBI, as shown in Appendix B. Ethical approval was received

covering the individual populations involved. This teaching data set consists of three clinical examinations: the Original Cohort, the Offspring Cohort, and the Third Generation Cohort. They are very suitable for an undergraduate or postgraduate bio-statistics research.

The design of data collection is central to any research. For a quantitative research, the process of measurement directly decides the outcome of the investigation as it accomplishes the connection between empirical evidence and quantitative expression (Merriam & Tisdell, 2016). The design of the FHS was conducted under the direction of the NHLBI and carefully monitored over three generations, which is trustworthy for us to do the CVD prediction research.

The data has been collected by lab assays, questionnaires and clinical tests approximately every two years since the start of the FHS. All participants were continuously followed through surveillance for outcomes of CVD regularly. Data of risk factors such as ECG, smoking history, blood pressure, medication etc. was gathered in each examination. All records must be checked and reviewed by professional physicians. Apart from these markers, validated events about cardiovascular related diseases (such as heart failure, stroke, cerebrovascular disease) were also recorded.

3.3.3 Data Extraction

According to Table 3.2, data collected in the first exam ("Exam 1") from the Framingham original cohort study include the maximum number of samples, 5209 subjects aging from 28 to 74. Considering the sample size, the data frame from "Exam 1" was chosen to develop the CVD prediction model, the one marked with the blue background in Table 3.2. The other five exams ranging from 8 to 12 (marked with the green background) will be used for the validation of the fitted model.

For "Exam 1", characteristics of 76 risk factors were gathered for each participant.

Data descriptions of complete columns are listed in Appendix C. According to the literature review on the risk factors in the CVD prediction models (see Section 2.3.3), some variables in the complete list will be removed. Finally, 20 candidate predictors will be included in the process of data analysis, listed in Table 3.3. Data frames used in the model validation (Exam 8, Exam 9, Exam 10, Exam 11, and Exam 12) will be extracted after the model has been developed and will be introduced in Chapter 5.

Table 3.3: Description of Candidate Predictors

ORDERS	PREDICTORS	UNITS	TYPES
1	AGE	YEARS	CONTINUOUS
2	SEX	0001 MALE , 0002 FEMALE	CATEGORICAL
3	BMI	KG/M2	CONTINUOUS
4	HYPERTENSION	0000 NEGATIVE, 0001 TRANSIENT, 0002 PERMANENT, 0003 TYPE UNKNOWN, 0008 DOUBTFUL	CATEGORICAL
5	HISTORY OF NERVOUS HEART	0000 NO, 0001 YES, DEFINITE	CATEGORICAL
6	HISTORY OF PERICARDITIS	0000 NO, 0001 YES, DEFINITE	CATEGORICAL
7	HISTORY OF OTHER CVD	0000 NO, 0001 YES, DEFINITE	CATEGORICAL
8	PREMATURE BEATS	0000 NO, 0001 YES, DEFINITE, 0002 YES, DOUBTFUL	CATEGORICAL
9	HISTORY OF ATRIOVENT RICULAR BLOCK	0000 NO, 0001 YES, DEFINITE, 0002 YES, DOUBTFUL	CATEGORICAL
10	HISTORY OF RHEUMATIC FEVER	0000 NONE, 0001 YES, 0008 DOUBTFUL	CATEGORICAL
11	HISTORY OF ALLERGY OR ASTHMA	0000 NEGATIVE, 0001 ALLERGY, ALONE, 0002 BRONCHIAL ASTHMA, ALONE, 0003 ALLERGY AND ASTHMA, TOGETHER	CATEGORICAL
12	HISTORY OF THYROID DISEASE	0000 NEGATIVE, 0001 HYPERTHYROID ONLY, 0002 HYPOTHYROID ONLY	CATEGORICAL
13	HISTORY OF SUBACUTE ENDOCARDITIS	0000 NO, 0001 YES	CATEGORICAL
14	BLOOD PRESSURE SYSTOLIC	MM HG	CONTINUOUS
15	BLOOD PRESSURE DIASTOLIC	MM HG	CONTINUOUS
16	CIGARETTES PER DAY	LAPSE, FORM 8/50	CONTINUOUS
17	CIGARS PER DAY	LAPSE, FORM 8/50	CONTINUOUS
18	PIPERS PER DAY	LAPSE, FORM 8/50	CONTINUOUS
19	PULSE RATE	PER MINUTE	CONTINUOUS
20	DIABETES	0000 NO, 0001 YES, DEFINITE	CATEGORICAL

Data of the candidate predictors are distributed in different files. If we manually extract data frames from these different files, it is inefficient, especially when we want to find candidate data for a specific subject, it needs to repeatedly query in multiple files. So we will write a piece of python code that helps us extract data automatically. The code flow chart is described in Figure 3.2.

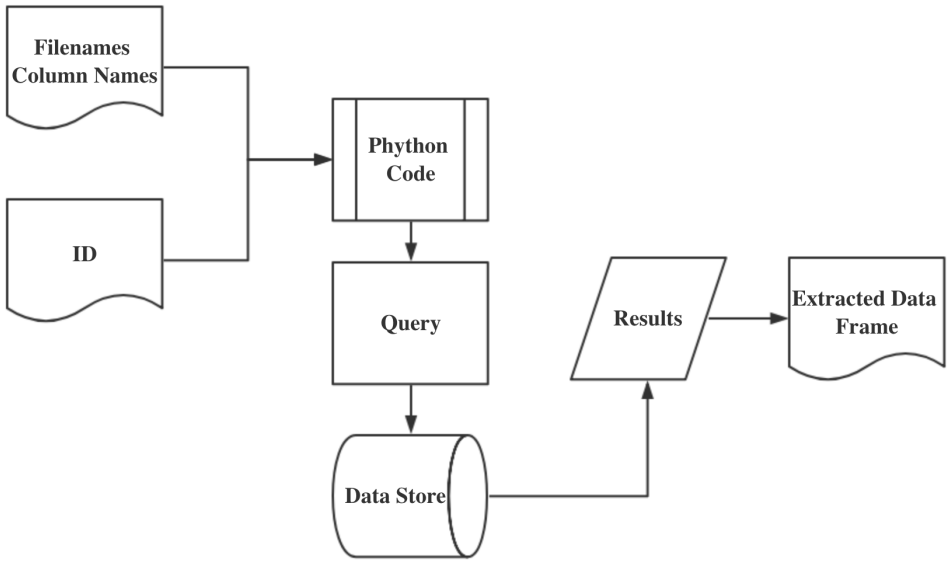


Figure 3.2: Flow Chart of Data Extraction Code

3.4 Data Analysis

3.4.1 Choice of the Modelling Method

The aim of this research is to develop a risk prediction model where multiple parameters are included. We want to estimate how various predictors simultaneously affect the probability of developing CVD for an individual. As was reviewed in Section 2, mainly three types of statistical methods can be used for survival analysis: parametric approaches, semi-parametric approaches, and non-parametric approaches (E. T. Lee & Wang, 2003).

The non-parametric approaches such as the Kaplan-Meier model (Kaplan & Meier, 1958) and the Log-rank test (Mantel, 1966; Cox, 1992) can only do univariate analysis with a single predictor, and they are not suitable for the analysis of continuous variables.

Both parametric approaches and semi-parametric approaches can do multiple parameter analysis. They all assume that the predictors and the log hazard rate should have a linear relationship between them (Efron, 1977); however, the Cox Proportional Hazard

model (the most popular method belonging to the semi-parametric statistical method) has the advantage that only the rank orderings of the failure and censoring times are used to estimate and test the regression coefficients (Cox, 1992). The Cox model is more efficient even though the assumption of the parametric models is met.

In addition, when the assumptions are not met, the Cox regression analysis can be still used efficiently with a extended Cox regression form (F. Harrell, 2013), but a parametric model such as Weibull survival distribution would be proved as a null model. Furthermore, methods for diagnosing the assumption of the Cox regression model are well developed. Thus, the Cox regression model was chosen as our statistical method for developing the CVD prediction model.

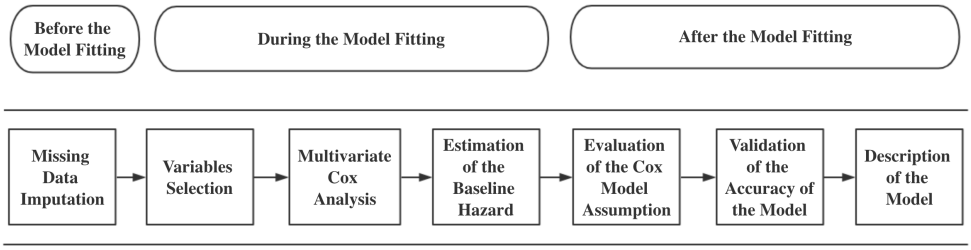


Figure 3.3: Procedures of the Cox Regression Analysis

3.4.2 Procedures of the Cox Regression Analysis

F. Harrell (2013) pointed out that several aspects should be considered when fitting multivariate prediction models. These aspects include processing of incomplete data (missing values), variable selections, model training and coefficient estimation, evaluation and validation of the fitted model, and presentation of the fitted model. This rule is suitable for any regression analysis. We applied this principle to the Cox regression analysis and did it in a step-wise manner referring to the characteristics of the Cox model (Cox & Oakes, 1984):

- Imputation of data with missing values
- Selection of variables
- Multiple Cox analysis
- Estimation of the baseline hazard rate
- Evaluation of the Cox model assumption
- Validation of the accuracy of the prediction
- Presentation of the fitted model

For a more intuitive expression, steps to conduct the Cox analysis are expressed using a flow chart, see Figure 3.3.

3.4.3 Tools and Packages Used

The R language (Team et al., 2013) was employed to compute and fit the model. Packages and functions that were used are listed in Table 3.4. More specific descriptions and uses of these packages and functions will be introduced in the corresponding sections.

Table 3.4: Summary of Tools and Packages

Packages	Functions	Functionalities
survival	coxph()	Therneau's function to fit the Cox model
	basehaz()	Computation of baseline hazard rate
	survfit()	Estimation of survival curves
	cox.zph()	Proportional hazard test
survminer	ggplot()	Visualisation of the results of analysis
rms	transcan()	Data imputation
	validate()	Model valiation
	calibrate()	Model calibration
	nomogram()	Nomogram of the model
	cph()	Modification and extension of coxph()

3.5 Summary

In summary, the core of this chapter was to establish a research method suitable for this research. We designed our research methodology referring approaches of research design by Creswell (2014), from the philosophical worldview to the specific research method. According to the features of this research, we decided to use the quantitative research method to analyse the data we collect. After that, more specific steps of conducting the research were stated, including the process of data collection, data extraction, data analysis, and the tools that will be employed.

Chapter 4

Cox Regression Prediction Model with Framingham Data Set

4.1 Introduction

In this chapter, we will demonstrate a full process of developing CVD risk prediction models using the Cox regression method. Sections will be organised as follows:

- Section 4.2 presents the theory of the Cox proportional hazard regression analysis, including the definition of the formula, as well as optional methods to estimate the regression coefficients.
- Section 4.3 introduces the process of imputing missing values in our data set, including identifying the type of missing value, choosing a proper imputation method, and the implementing the process of data the imputation.
- Section 4.4 tells the process of variable selection from candidates risk factors according to the statistical outputs derived.
- Section 4.5 states the process of multiple variable analysis using the Cox regression model. Statistics will be generated and interpreted and a CVD risk estimation

model will be developed.

- Section 4.6 illustrates how we derived the baseline hazard rate in the Cox formula.
- Section 4.7 presents the theory of the proportional hazard assumption of the Cox model, as well as the process of checking this assumption.

4.2 Theory of Cox Regression Model

4.2.1 Definition of the Cox Regression Model

The Cox regression Model is a Proportional Hazards (PH) technique commonly used for investigating the association between a few explanatory variables and the survival time of a subject (Cox, 1992). The feature of the Cox regression model is that it can investigate how several factors simultaneously affect the occurrence of the event. In survival analysis, the Cox model allows us to examine the rate of a specific event happening, e.g., the development of CVD.

The predictor variables are generally described as covariates in the Cox formula. The Cox model is expressed by a hazard function and a set of predictor variables according to time t. The t in the Cox model indicates that the hazard will probably change over time. The Cox regression formula has the form:

$$\lambda(t;x) = \lambda(t)exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_m) \tag{4.1}$$

where,

- t represents the time that the event occurs.
- $\lambda(t;x)$ is the hazard function for a subject at time t. It is determined by a set of m covariates $(x_1, x_2, ..., x_m)$.
- x_1,x_2,...,x_m are the values of covariates for a subject.

- β_1,β_2,...,β_m are the regression coefficients that measure the effect size of covariates.
- exp is the exponential function ($exp(x)$= e^x).
- $\lambda(t)$ is the baseline hazard rate, an arbitrary (unknown) function, corresponds to the hazard value when all x_i equal to zero, which means that the value $exp(0)$ equals 1.

The Cox regression formula can be presented as a multi-linear regression by taking the logarithm of the hazard $\lambda(t;x)$ on covariates x_i. The baseline hazard rate is an 'intercept' that changes over time, see the equation 4.2.

$$log(\lambda(t;x)) = log(\lambda(t)) + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_m \tag{4.2}$$

Another form of the Cox regression model is equation 4.3. The $exp(\beta_i x_i)$ is called hazard ratios (HR). A HR above one suggests that the hazard of developing an event will increase, i.e. the length of survival will decrease, along with the value of the i_{th} covariate increases. In the other words, it means that a predictor variable positively has effect on the probability of developing the event, i.e. negatively associates with the length of survival.

$$HR = \frac{\lambda(t;x)}{\lambda(t)} = exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k) \tag{4.3}$$

When,

- HR = 1: indicates that there is no effect on the hazard
- HR < 1: indicates a reduction in the hazard
- HR > 1: indicates an increase in the hazard

4.2.2 Estimation of Regression Coefficients

Let $t_1 < t_2 < ... < t_m$ be the failure times. The index m is unique and does not include censored observations in the sample of n subjects. The set of failures (deaths) in an instant before failure time t_i is referred to R_i, which is called the *risk set* at time t_i. R_i includes the set of subjects in which the subjects j had not been censored or failed by time t_i. In other words, subjects with censoring or failing time $Y_j \geq t_i$ are included in this set of R_i. The regression coefficients of the cox model can be computed based on the partial likelihood method (Cox, 1992).

4.2.2.1 The Log Likelihood Estimation

For now, when no ties among the failure times, the $m = n$. Given that the individuals in the risk set R_i are failing and one failure exactly occurs at time t_i, then by applying rules of conditional probability, the conditional probability that this subject i who failed at t_i is

$$\text{Prob}\{\text{subject } i \text{ fails at } t_i | R_i \text{ and one failure at } t_i\} = \frac{\text{Prob}\{\text{subject } i \text{ fails at } t_i | R_i\}}{\text{Prob}\{\text{one failure at } t_i | R_i\}} \tag{4.4}$$

The conditional probability in formula 4.4 can be expressed as independent of λ(t), as shown in formula 4.5 (F. Harrell, 2013).

$$\frac{\lambda(t_i)exp(x_i\beta)}{\sum_{j \in R_i} \lambda(t_i)exp(x_i\beta)} = \frac{exp(x_i\beta)}{\sum_{j \in R_i} exp(x_i\beta)} = \frac{exp(x_i\beta)}{\sum_{Y_j \geq t_i} exp(x_j\beta)} \tag{4.5}$$

When predictors have no effect on the likelihood, i.e. β= 0, then exp(x_iβ) = 1 and exp(x_jβ) = 1. This likelihood described in formula 4.5 equals to $1/n_i$, and n_i represents the number of risk individuals at time t_i.

Cox (1992) argued that the conditional probabilities that fail across different failure times are not conditionally dependent, then by multiplying the likelihood for each individual over all failure times, a total likelihood (also called *the partial likelihood* for β) can be estimated, as shown in formula 4.6 .

$$L(\beta) = \prod_{Y_i \ uncensored} \frac{exp(x_i\beta)}{\sum_{Y_j \geq Y_i} exp(x_j\beta)} \tag{4.6}$$

Correspondingly, the log partial likelihood can be obtained as:

$$logL(\beta) = \sum_{Y_i \ uncensored} \{x_i\beta - log[\sum_{Y_j \geq Y_i} exp(x_j\beta)]\} \tag{4.7}$$

It is proven that Maximum Likelihood Estimations (MLEs) of β deprived from this partial log-likelihood are valid (Cox, 1992). This log-likelihood will not be affected when adding a constant to any or all of the covariates.

4.2.2.2 Breslow's Estimation

When having tied survival times in the set of subjects, the true partial likelihood becomes very cumbersome. Computation of this exact likelihood involves permutations and consumes time. Breslow (Breslow, 1974) promoted an approximate log-likelihood function which satisfies this situation as there are tied failure times among the samples so that $m < n$. The m still indexes the unique order of failure times and the risk set is denoted as $t_1 < t_2 < ... < t_m$. Let d_i represent the number of subjects failing at time t_i. Using Breslow's estimation, the log likelihood deprived above is written as

$$logL(\beta) = \sum_{i=1}^{k} \{\sum_{j \epsilon D_i} x_j\beta \ - \ d_i log[\sum_{Y_j \geq t_i} exp(x_j\beta)]\} \tag{4.8}$$

Let $S_i = sum_{j \epsilon D_i} x_j$, where D_i represents the set of subjects indexed by j at time t_i, then the Breslow's estimation of log likelihood can be simplified as

$$logL(\beta) = \sum_{i=1}^{k}\{S_i\beta \; - \; d_i log[\sum_{Y_j \geq t_i} exp(x_j\beta)]\} \tag{4.9}$$

4.2.2.3 Efron's Estimation

Apart from the exact log likelihood and Breslow's approximation to the log likelihood, Efron (Efron, 1977) did another deprivation on the approximation to the true likelihood for censored data. This estimation is very close to the cumbersome permutation likelihood so Efron's estimation is regarded as more accurate than the Breslow's (He & Zaslavsky, 2012). Efron's estimation is written as

$$\begin{aligned} logL(\beta) = \sum_{i=1}^{k}\{S_i\beta \; - \; \sum_{j=1}^{d_i} d_i log[\sum_{Y_j \geq t_i} exp(x_j\beta)] \\ - \; \frac{j \; - \; 1}{d_i} \sum_{l \epsilon D_i} exp(x_l\beta)\} \end{aligned} \tag{4.10}$$

4.3 Imputation for Missing Data

4.3.1 What is Missing Data

Missing data is a widespread problem that an analyst likely encounters in data analysis, particularly in health or clinical research where complete data sets are rare. There are various reasons of missing data. In some cases, data is missing randomly but sometimes the participants intentionally conceal some of their information. Incomplete data might cause bias in predictions and influence the validity of research results (Rubin, 1996). Considering the issue of missing data prior to modelling is important.

Three types of missing data can be defined according to reasons that the missing values occurred, namely Missing Completely At Random (MCAR), Missing At Random (MAR) and Informative Missing (IM) (Enders, 2010). MCAR occurs when data are randomly missing unrelated to any reason. Compared with MCAR, MAR and IM are the

situations where data are not missing at random. MAR relates to some characteristics or the response of the subjects where some information is intentionally concealed. IM is the situation where data are missing with true values very informative, which means that the missing data are important and cannot be ignored.

4.3.2 Missing Data in the FHS Data Set

Table 4.1 lists the missing values for each variable in the data frame selected in section 3.3.3, see the column "NAs" (Not Available). According to this table, variables that do not have missing data include "AGE", "SEX", "HISTORY OF ATRIOVENT RISCULAR BLOCK", "HISTORY OF ALLERGY OR ASTHMA", "PREMATURE BEATS" and "PULSE RATE". However, the remaining variables all have missing values.

Table 4.1: Missing Values of Candidate Predictors

Predictors	Variables	NAs
AGE	age	0
SEX	sex	0
BMI	bmi	10
HYPERTENSION	hyp	5
HISTORY OF NERVOUS HEART	honh	136
HISTORY OF PERICARDITIS	hop	136
HISTORY OF OTHER CVD	hooc	136
HISTORY OF ATRIOVENT RICULAR BLOCK	pb	0
HISTORY OF RHEUMATIC FEVER	hoarb	14
HISTORY OF ALLERGY OR ASTHMA	horf	0
HISTORY OF THYROID DISEASE	hoaoa	25
HISTORY OF SUBACUTE ENDOCARDITIS	hotd	136
PREMATURE BEATS	hose	0
BLOOD PRESSURE SYSTOLIC	bps	1952
BLOOD PRESSURE DIASTOLIC	bpd	1952
CIGARETTES PER DAY	cgrpd	2154
CIGARS PER DAY	cgpd	2161
PIPERS PER DAY	ppd	2160
PULSE RATE	pr	0
DIABETES	dia	1180

Provided that the types of missing values can be known or estimated, appropriate methods can be applied to deal with the missing data but unfortunately most of them cannot be guaranteed. Most imputation methods assume the type of missing data is MAR. In this research, let's assume the type of missing values in our data set is MAR. The simulation data are approximate to the true values as long as enough variables are added into the imputation algorithms (F. Harrell, 2013).

4.3.3 Methods of Imputation

As discussed above, missing data occurs frequently, and the prelude to developing a model is to handle these missing data. There are many imputation approaches available. The simple method to deal with the missing data is just deleting the incomplete cases, which is the so-called "complete case method" (Enders, 2010). However, this method is only unbiased when the type of incomplete data is missing as MCAR, which is rare in actuality (Kenward, 2013). In addition, if there are many missing values in the data set, a large proportion of sample cases will be dropped off. The size of the remaining samples may not be large enough to result in a model with good performance when this "complete case method" is applied (F. Harrell, 2013).

As shown in Table 4.1, 14 of 20 variables have missing values. For example, the number of missing values for "CIGARETTES PER DAY" is 2154, and as the number of complete cases is 5079, approximately half of the cases are incomplete. Provided the "complete case method" is applied, i.e. directly removing the cases with missing values, there will be a large number of cases deleted, which will seriously decrease the size of the sample. Thus, in this research, we need to choose a suitable imputation method for mocking the missing data.

The alternative approach to handling the incomplete data is to replace the missing

data with substituted values (Rubin, 1996). Available approaches for generating substituted values are: single imputation and multiple imputation (Sterne et al., 2009). *Single imputation* uses techniques such as the maximum likelihood, mean substitution, weighting methods, Bayesian inference, or others to estimate the substituted values just once (Horton & Kleinman, 2007). Compared with single imputation, in multiple imputation missing values are imputed using similar techniques applied in single imputation but is more complicated. *Multiple imputation* involves a much more sophisticated process of creating the imputation model using statistical techniques, simulating the best guess data, result analysis, and then pooling (Van Buuren, 2012).

Multiple imputation has proven a much more valuable method to approach missing data due to its valid inference process mentioned above. Another good feature of multiple imputation over single imputation is that it is more flexible. It can be used in different scenarios no matter what type of data is missing (Royston et al., 2004; White, Royston & Wood, 2011). That is why multiple imputation has become one of the leading methods for data simulation.

As reviewed above, the multiple imputation method is one popular technique to generate substituted values. We will use multiple imputation to manage the missing data in our data set.

4.3.4 Implementation of the Imputation

The R function $transcan$ in $Hmisc$ package was used. $Transcan$ can do data transformation and multiple imputation. This function transforms and imputes continuous and categorical variables using algorithms simply modified from Maximum Generalised Variance (MGV) in the SAS PRINQUAL procedure (Kuhfeld, 1990). This SAS procedure makes sure that imputed values have a maximum correlation with the best linear combination of other variables. Expected "best guess" values are simulated for

replacing the missing values in the data set.

Before we start imputing the missing data, we assign each candidate predictor a variable name listed in Table 4.1, the column "variables". R code of the imputation is described in Code Snippet 4.1. Variables with missing values (bmi, bps, hyp, honh, hop, hooc, hose, bpd, ppd, cgpd, cprpd, and dia) in Table 4.1 are parameters passed into the $transcan$ function. pr was set to "FALSE" for not printing R^2 and shrinkage factors. pl was set to "FALSE" for not plotting the distribution of imputed values.

```
w.transcan <- transcan(~ bmi + bps + hyp + honh + hop
                       + hooc + hose + bpd+ ppd + cgpd
                       + cgrpd + dia, imputed=TRUE, trantab=TRUE,
                       data=FOCExam1Data, pl=FALSE, pr=FALSE)
# summarise the imputation result
summary(w.transcan, long=T)
```

Code Snippet 4.1: R Code for Imputation

The result of the imputation is presented in Table 4.2. The imputed values of missing data for each variable are shown in column "Num. of NAs Imputed". The numbers should be consistent with "NAs" in Table 4.1. The column "Missing" for each variable in Table 4.2 is zero, which indicates all missing data are imputed. "Mean of Imputed Values" is also calculated. For example, the number of missing values imputed for the predictor "BMI" is 10, which is consistent with the number presented in column "NAs" of Table 4.1. No missing values are left out (the value of "Missing" is 0), and the mean of imputed missing values is 23.04.

4.4 Variable Selection

We have a list of candidate predictors shown in Table 4.1 but probably not all of them are significant to the development of CVD. Appropriate variables should be pre-specified before we start to do the multiple regression analysis. The goal of variable selection is

Table 4.2: Results of Imputation

Predictors	Variables	Num. of NAs Imputed	Missing	Mean of Imputed Values
AGE	age	/	/	/
SEX	sex	/	/	/
BMI	bmi	10	0	23.04
HYPERTENSION	hyp	5	0	0.03671
HISTORY OF NERVOUS HEART	honh	136	0	0.009014
HISTORY OF PERICARDITIS	hop	136	0	0.001563
HISTORY OF OTHER CVD	hooc	136	0	0.04275
HISTORY OF ATRIOVENT RICULAR BLOCK	pb	/	/	/
HISTORY OF RHEUMATIC FEVER	hoarb	14	0	0
HISTORY OF ALLERGY OR ASTHMA	horf	/	/	/
HISTORY OF THYROID DISEASE	hoaoa	25	0	7.687
HISTORY OF SUBACUTE ENDOCARDITIS	hotd	136	0	0.0008399
PREMATURE BEATS	hose	/	/	/
BLOOD PRESSURE SYSTOLIC	bps	1952	0	139.9
BLOOD PRESSURE DIASTOLIC	bpd	1952	0	79.85
CIGARETTES PER DAY	cgrpd	2154	0	21.35
CIGARS PER DAY	cgpd	2161	0	2.937
PIPERS PER DAY	ppd	2160	0	0.8969
PULSE RATE	pr	/	/	/
DIABETES	dia	1180	0	0.01727
Notes: '/' indicates imputation is not appliable because of no missing values				

to identify prognostic factors that are independently significant for developing CVD.

There are multiple options for variable selection (see Figure 4.1). The following sections demonstrate why we chose "Forward Selection".

4.4.1 Methods for Variable Selection

Two methods are available for the variable selection: univariable screening (Hammermeister, DeRouen & Dodge, 1979) and stepwise variable selection (Kano & Harada, 2000). Compared with univariable screening, stepwise variable selection is much more popular as prognostic factors will be identified step by step (Kano & Harada, 2000). We decided to use stepwise variable selection as our method for factor analysis.

When employing stepwise variable selection, three methods within it can be applied:

Partial Least Squares (PLS) regression, Least absolute shrinkage and selection operator (Lasso), and Stepwise Regression (Chong & Jun, 2005). Among these methods, Stepwise Regression is regarded as a standard procedure for selecting factors. Objective statistics are generated to help us to do the variable selection and then predictors will be introduced into the model sequentially.

There are three options in Stepwise Regression: forward selection, backward elimination and stepwise method (Chong & Jun, 2005). In forward selection, one predictor is added into the regression model at a time, while the backward elimination adds all predictors at first and then eliminates the insignificant ones successively. The stepwise method starts with forward selection but considers the backward elimination when deleting a predictor at each stage. These three methods are expected to have similar performance (Chong & Jun, 2005). The hierarchy of variable selection methods is demonstrated in Figure 4.1. This graph also shows the route of selecting "Forward Selection" as our method for filtering redundant variables.

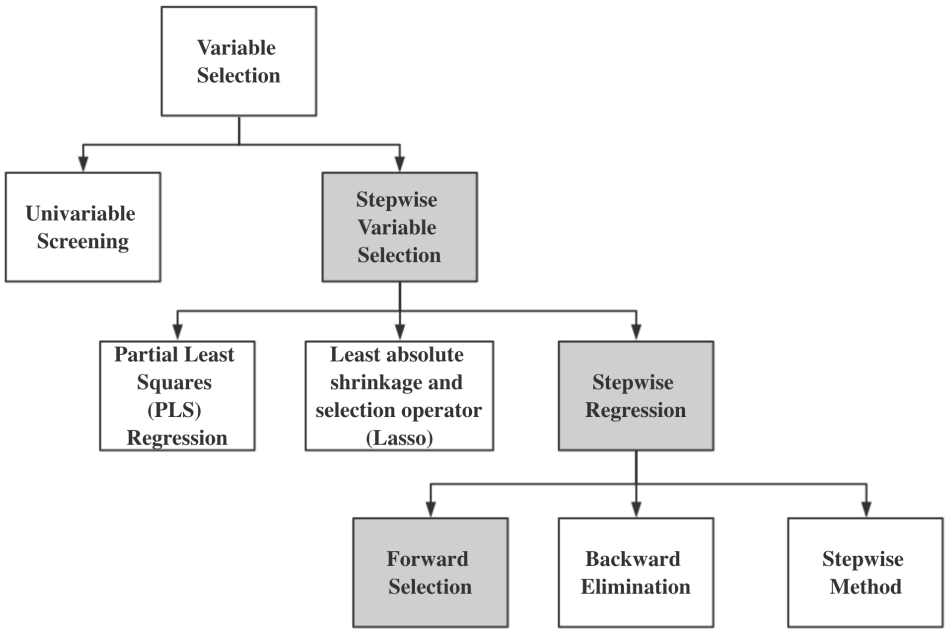


Figure 4.1: Summary of Optional Variable Selection Methods

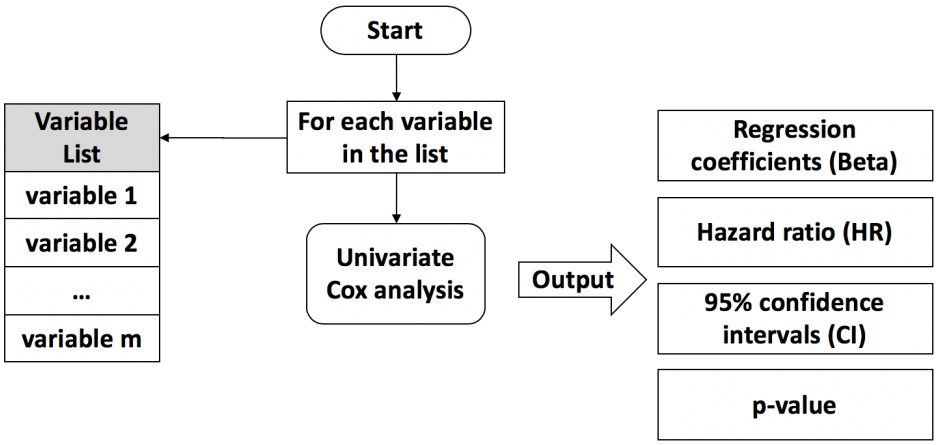


Figure 4.2: Interpretation of the R Code for Forward Variable Selection

4.4.2 Implementation of Variable Selection

According to the rules in the Forward Selection method, we will implement the process of variable selection following two steps:

- Step 1: applying the univariate Cox analysis for each candidate variable
- Step 2: filtering variables that are not significant to the development of CVD according to a criterion we defined

In step 1, candidate predictors in Table 4.1 are independently inputted to the Cox regression model one by one (also called univariate Cox analysis). R code listed in Code Snippet D.1 of Appendix D accomplishes applying the univariate Cox analysis to multiple variables at once. Figure 4.2 presents a graphical explanation of the R code for univariate Cox analysis.

Table 4.3 reports the statistical form values of the forward variable selection for assessing the significance of each factor. The outputs for each of the variables include the regression beta coefficients (given as "beta"), the effect sizes with lower .95 and upper .95 confidence interval (given as HR) and the statistical significance (given as p.value) in relation to overall survival. More specific interpretations are listed below:

Table 4.3: Statistical Outputs of Forward Variable Selection

Predictors	Variables	Beta	HR (95% CI for HR)	P.value	Signif. Codes
AGE	age	0.061	1.1 (1.1-1.1)	<2e-16	***
SEX	sex	-0.49	0.62 (0.57-0.66)	<2e-16	***
BMI	bmi	0.062	1.1 (1.1-1.1)	<2e-16	***
HYPERTENSION	hyp	0.18	1.2 (1.2-1.2)	<2e-16	***
HISTORY OF NERVOUS HEART	honh	0.041	1 (0.72-1.5)	0.83	
HISTORY OF PERICARDITIS	hop	1.1	2.9 (1.4-6)	0.0055	.
HISTORY OF OTHER CVD	hooc	0.48	1.6 (1.4-1.9)	4.00E-08	***
HISTORY OF ATRIOVENT RICULAR BLOCK	pb	0.067	1.1 (0.92-1.2)	0.38	
HISTORY OF RHEUMATIC FEVER	hoarb	0.0208	1.02(0.99-1.05)	0.145	
HISTORY OF ALLERGY OR ASTHMA	horf	-0.0064	0.99(0.92-1.07	0.861	
HISTORY OF THYROID DISEASE	hoaoa	-0.0145	0.98(0.95-1.02)	0.42	
HISTORY OF SUBACUTE ENDOCARDITIS	hotd	0.8754	2.39(0.77-7.44)	0.13	
PREMATURE BEATS	hose	0.13	1.1 (1-1.3)	0.01	*
BLOOD PRESSURE SYSTOLIC	bps	0.022	1 (1-1)	<2e-16	***
BLOOD PRESSURE DIASTOLIC	bpd	0.037	1 (1-1)	<2e-16	***
CIGARETTES PER DAY	cgrpd	0.0091	1 (1-1)	5.80E-06	***
CIGARS PER DAY	cgpd	0.038	1 (1-1.1)	0.0029	*
PIPERS PER DAY	ppd	0.031	1 (0.99-1.1)	0.098	.
PULSE RATE	pr	0.0059	1.005886	0.00115	**
DIABETES	dia	1.5	4.7 (3.7-6)	<2e-16	***
Note: Signif. codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**					

- The column Beta lists the regression coefficients. A positive sign of a variable indicates that the risk (hazard) of developing CVDs is higher for subjects with higher values, and thus worse prognosis. For example, the predictor variable "sex" has two values: 1: male, 2: female. The regression coefficient (see the column "Beta") for sex is -0.49, which means that women have a lower risk of developing CVD than men.
- The column HR is the hazard ratio, which is the exponential coefficient. It gives the impact of covariates. For example, the HR for sex is exp(coef) = exp(-0.49) = 0.62. This means that being a woman reduces the risk of developing CVD by a factor of 0.62, which is associated with a good prognosis compared to men.
- The 95% CI for HR is the confidence intervals (CI) for hazard ratios. This gives us the upper 95% CI and lower 95% CI for HR. For the variable sex, the lower 95% bound is 0.57, the upper 95% bound is 0.66.
- The last feature in the output is the p-value that represents the global statistical

significance of the fitted model.

In step 2, insignificant variables need to be removed. The criteria of adding or removing a variable from a regression model is the significance level p-value. This p-value can range from 0.05, 0.1, 0.15, and 0.2 based on multiple sequential hypothesis testing of individual variables (Allen, 1974; Tartakovsky, Nikiforov & Basseville, 2014). A p-value less than 0.05 is defined as a strict inclusion criteria for variable selection (Altman & Royston, 2000). In our work, we accept "p-value" < 0.5 as our inclusion criterion. According to the inclusion criteria defined above, covariates regarded as having significance to the final risk model should have p-value <0.05. Figure 4.3 graphically demonstrates the variable inclusion criterion as well as the whole process of filtering insignificant variables in step 2. Variables with p-value <0.5 were included but these variables with p-value higher than 0.05 were removed.

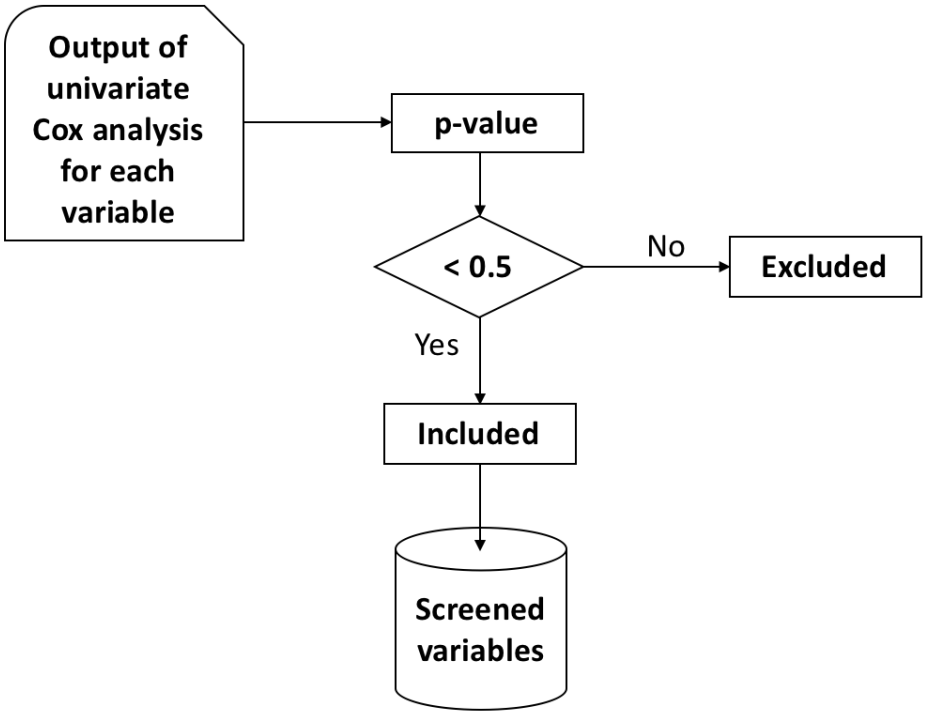


Figure 4.3: Procedures for Variable Selection Using Forward Selection

Lastly, candidate prognostic variables that mostly determine predictive of survival

are screened, as shown in Table 4.4.

Table 4.4: Results of Variable Selection

Rank	Predictors	Variables
1	AGE	age
2	SEX	sex
3	BMI	bmi
4	HYPERTENSION	hyp
5	HISTORY OF PERICARDITIS	hop
6	HISTORY OF OTHER CVD	hooc
7	PREMATURE BEATS	hose
8	BLOOD PRESSURE SYSTOLIC	bps
9	BLOOD PRESSURE DIASTOLIC	bpd
10	CIGARETTES PER DAY	cgrpd
11	CIGARS PER DAY	cgpd
12	PULSE RATE	pr
13	DIABETES	dia

4.5 Multivariable Analysis

Multivariable analysis is a way using statistical techniques to determine how different causes relatively contribute to a single outcome (Katz, 2011). The goal of the multivariable analysis in this paper is to see how the candidate factors (see Table 4.4) jointly impact on the incidence of CVD in the Cox model. Variables identified as independently predictive of risk of the development of CVD are entered into the multivariate analysis. R code for the multivariable analysis is listed below.

```
1  res.cox.one <- coxph(Surv(cvddate, cvd) ~ age + sex + bmi
2                       + hyp + hop + hooc + pb + bps + bpd
3                       + cgrpd + cgpd + pr + dia,
4                       data =  FOCExam1Data)
5
6  summary(res.cox.one)
```

Code Snippet 4.2: R Code of Multivariable Analysis

Table 4.5: Statistical Outputs of Multivariable Analysis

Predictors	coef	exp(coef)	se(coef)	z	lower 0.95	upper 0.95	Pr(>|z|)	Signif. Codes
AGE	0.054923	1.056459	0.00426	12.893	1.0477	1.0653	2.00E-16	***
SEX	-0.493051	0.61076	0.071385	-6.907	0.531	0.7025	4.95E-12	***
BMI	0.017167	1.017315	0.008294	2.07	1.0009	1.034	0.038465	*
HYPERTENSION	0.189047	1.208097	0.100104	1.889	0.9929	1.47	0.058959	.
HISTORY OF PERICARDITIS	1.138342	3.121589	0.709373	1.605	0.7772	12.537	0.108556	
HISTORY OF OTHER CVD	0.414877	1.514184	0.242307	1.712	0.9417	2.4346	0.086861	.
PREMATURE BEATS	0.020743	1.020959	0.098749	0.21	0.8413	1.239	0.833625	
BLOOD PRESSURE SYSTOLIC	0.009667	1.009714	0.002492	3.879	1.0048	1.0147	0.000105	***
BLOOD PRESSURE DIASTOLIC	0.009496	1.009542	0.004435	2.141	1.0008	1.0184	0.08224	.
CIGARETTES PER DAY	0.010117	1.010168	0.002708	3.736	1.0048	1.0155	0.000187	***
CIGARS PER DAY	-0.017347	0.982802	0.017626	-0.984	0.9494	1.0173	0.325036	
PULSE RATE	0.003016	1.003021	0.002506	1.204	0.9981	1.008	0.228749	
DIABETES	1.571486	4.813794	0.212859	7.383	3.1718	7.3059	1.55E-13	***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1								

Table 4.5 lists the statistical output of the multivariate Cox regression analysis. Interpretations of the output are:

- coef: regression coefficient,
- exp(coef): exponentiated coefficients: HR,
- se(coef): standard error of coefficient,
- z: equals to coef/se(coef), which is the statistical significance and displays the wald statistic value as default,
- lower .95 & upper .95: the upper and lower confidence intervals
- Pr(>|z|): the statistical significance of each variable to the model

Observing the statistics in Table 4.5, there are still large p-values for some variables. Insignificant predictors need to be filtered again for the final model. Applying the same inclusion criteria as in the process of the forward variable selection, predictors with a p-value greater than 0.05 fail to be significant and will be removed. Table 4.6 lists the predictors that are finally selected to develop the CVD prediction model.

A special case is the predictor "PULSE RATE". A great many of articles have been researching the relationship between heart rate and CVD. A strong association between heart rate variability and cardiovascular-related disease has been proven (Han et al., 2017; Yu et al., 2017; Böhm et al., 2010). Even though "PULSE RATE" has a statistical

Table 4.6: Final Variables Entering the Risk Model

Rank	Predictors	lower 0.95	upper 0.95	Pr(>|z|)	Signif. Codes
1	AGE	1.0477	1.0653	2.00E-16	***
2	SEX	0.531	0.7025	4.95E-12	***
3	BMI	1.0009	1.034	0.038465	*
4	HYPERTENSION	0.9929	1.47	0.058959	.
5	BLOOD PRESSURE SYSTOLIC	1.0048	1.0147	0.000105	***
6	CIGARETTES PER DAY	1.0048	1.0155	0.000187	***
7	PULSE RATE	0.9981	1.008	0.228749	
8	DIABETES	3.1718	7.3059	1.55E-13	***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Table 4.7: Statistical Outputs of Multivariable Cox Analysis Using Final Variables

Predictors	coef	exp(coef)	se(coef)	lower .95	upper .95	z	Pr(>|z|)
log(AGE)	2.083643	8.03369	0.11535	6.4082	10.0716	18.064	< 2e-16
SEX	-0.46972	0.62518	0.0394	0.5787	0.6754	-11.922	< 2e-16
log(BMI)	0.608864	1.83834	0.12575	1.4368	2.3521	4.842	1.29E-06
HYPERTENSION	0.241461	1.27311	0.05894	1.1342	1.429	4.097	4.19E-05
log(BLOOD PRESSURE SYSTOLIC)	1.682571	5.37937	0.17817	3.7938	7.6277	9.443	< 2e-16
CIGARETTES PER DAY	0.009669	1.00972	0.00165	1.0065	1.013	5.857	4.70E-09
log(PULSE RATE)	-0.30209	0.73927	0.11692	0.5879	0.9297	-2.584	0.00978
DIABETES	1.087501	2.96685	0.12451	2.3244	3.7869	8.734	< 2e-16
Concordance = 0.706 (se = 0.006)							
Likelihood ratio test. = 1330 on 8 df, p=0							
Wald test = 1326 on 8 df, p=0							
Score (logrank) test = 1417 on 8 df, p=0							

p-value of 0.228749 (higher than 0.05), we still include it in our final model analysis.

For estimating the maximum likelihood, three methods (the log-likelihood (the exact), Breslow's approximation to the log-likelihood and the Efron's approximation to the log-likelihood) stated in section 4.2.2 are available. Considering the ties handling and the computation consumed, Efron's log likelihood is chosen as the derivation of an estimator of β.

Now, all variables in Table 4.6 are added to the Cox regression model from which a predictive model is developed. As the Cox model is an exponential function with a vector of multiple predictors, all continuous variables will be taken as natural logarithms for making them an approximate linear distribution. The final outputs of the multiple variable Cox regression analysis are presented in Table 4.7.

Except for the statistics (the coef, exp(coef), se(coef), z, 95% CI, Pr(>|z|)) illustrated above, the final outputs in Table 4.7 also give the global statistical significance of the fitted model and the concordance. A high statistically significant relation of risk factors and the occurrence of CVD are observed according to the Pr(>|z|)(p-value).

The bottom row in Table 4.7 is the global statistical significance p-value using three tests: likelihood test, wald test and score (log-rank) test. The three tests soundly reject the omnibus null hypothesis, all coefficient betas (β) are 0. That means that the fitted model is significant (all global p-values are approximately equal to zero). The meaning of the concordance of the model will be explained in the next chapter.

The HRs of covariates are interpreted as having effects on the risk of getting CVD. For example, being a woman (sex = 2) reduces the hazard by a factor of 0.62 when holding the other covariates constant. Similarly, a person with diabetes has a hazard ratio HR(exp(coef)) = 4.71 that increases the risk of developing CVD. We can use the regression coefficients estimated from the Cox regression analysis and values of covariates for a specific individual to compute his probability of developing CVD.

4.6 Estimation of the Baseline Hazard Function

So far, the model has been developed and the vector of regression coefficients $(\beta_1, \beta_2, ..., \beta_m)$ has been derived. The left-hand side of the formula Cox (see Equation 4.1), i.e. $\lambda(t)$, should be estimated for further computation of an individual's probability of developing CVD. Cox (1992) assumed that the baseline hazard can take any shape, but it cannot be negative. It is estimated nonparametrically.

An R function $basehaz$ was used to estimate the baseline hazard rate. This function can compute the baseline hazard based on all covariates equal to zero or at mean values. R code for computing the baseline hazard is listed in Code Snippet 4.3.

```
# get baseline hazard with mean values of all covariates
b.mean <- basehaz(res.cph.final, centered = TRUE)
plot(b.mean$hazard)
b.mean$time
index = which(b.mean$time == 3655)
b.mean$hazard[index]

# get baseline hazard when all covariates equal to zero
b.zero <- basehaz(res.cph.final, centered = FALSE)
index = which(b.zero$time == 3655)
b.zero$hazard[index]
plot(b.zero$hazard)
```

Code Snippet 4.3: R Code of Computing the Baseline Hazard Rate

Table 4.8 lists values of the baseline hazard where the time point is 10 years. The 10-year baseline hazard rate equals to 0.1023354 at mean values of all covariates, 0.001863652 at all covariates equal to zero. Corresponding, the survival probabilities ($exp(-basehaz)$) were also calculated, which are 0.9027267 at mean values and 0.9981381 at all covariates equal to zero.

Table 4.8: Baseline Hazard and Survival Rate at 10 Years

	Covariates at mean value	**Covariates equal to zero**
Baseline hazard estimate	0.1023354	0.001863652
Baseline survival estimate	0.9027267	0.9981381

We will use the baseline hazard values as a building block for further calculations of the risk of developing CVD (or the survival probability) for an individual.

4.7 Evaluation of the Cox Model Assumption

4.7.1 The Assumption of the Cox Model

According to the Cox formula (Equation 4.1), two quantities produce the hazard at time t. The left part is the baseline hazard and the right part is the exponential expression e

to the sum of the m predictor variables X. An important characteristic of this formula is that the PH assumption should hold.

The Cox regression model assumes that the ratio of the hazards $\lambda(t;x)$ comparing any two collections of predictor variables $x_i = (x_{i1}, x_{i2}, ..., x_{ik})$ and $x_j = (x_{j1}, x_{j2}, ..., x_{jk})$ is constant over time t, $\lambda(t;x_i)/\lambda(t;x_j)$ (shown in Equation 4.11), does not depend on time t. The effect of a given predictor variable does not change over time.

$$\begin{aligned} HR = \frac{\lambda(t;x_i)}{\lambda(t;x_j)} &= \frac{\lambda(t)exp(\beta_1 x_{i1} + ... + \beta_k x_{ik})}{\lambda(t)exp(\beta_1 x_{j1} + ... + \beta_k x_{jk})} \\ &= exp\beta_1(x_{i1} - x_{j1}) + ... + \beta_p(x_{ik} - x_{jk}) \end{aligned} \tag{4.11}$$

For example, when considering 'sex' as a predictor to estimate CVD risk, if the risk of developing CVD for an individual (male,x = 0) is twice compared with another individual (female,x = 1) at age 45, the hazard ratio is e^{β}. This ratio is not depending on time t, then the risk will not change when they are at age 55 or any other age, as shown in the inferential process of Formula. 4.12.

$$HR = \frac{\lambda(t;x=1)}{\lambda(t;x=0)} = \frac{\lambda(t)e^{\beta} * 1}{\lambda(t)e^{\beta} * 0} = \frac{e^{\beta}}{e^{0}} = e^{\beta} \tag{4.12}$$

As the Cox regression model assumes that the effect of a time-independent variable is constant over time, which is also a restriction to use this model, it is important to verify whether variables included in a model satisfy the PH assumption.

4.7.2 Checking the PH Assumption

Generally, there are three approaches available to assess whether the hazards between two values of variables are proportional or not, as listed below. Each method has its advantages and disadvantages. Neither is regarded as the superior one.

- Graphical approaches: two types of graphical techniques are commonly used. One popular approach is to plot the "log-log" survival curves. The other is to compare the observed survival curves with the predicted curves.

- Goodness-of-fit (GOF) approaches: by employing statistical techniques, the chi-square and p-value for each predictive variable in the fitted model is computed to evaluate the PH assumption.

- Time-dependent variable approach: using this approach, the Cox model is extended to contain time-independent variables for assessing the PH assumption.

```
# check the PH assumption
test.coxph.final <- cox.zph(res.coxph.final)
test.coxph.final

ggcoxzph(test.coxph.final,
         ggtheme = theme_light(), point.col = "black",
         point.size = 1,point.alpha = 0.5)
```

Code Snippet 4.4: R Code of Checking PH Assumption

The GOF approach was used for checking the PH assumption. The biggest appealing advantage of the GOF approach compared with the other two techniques is that a statistical test value (p-value) for a given variable will be provided, so the researcher can make a clear-cut decision according to the test statistics. Schoenfeld residuals are used (Schoenfeld, 1982) to diagnose the PH assumption. Correlations between ranked failure times and Schoenfeld's residuals for each predictor will be obtained using a chi-square statistic with 1 df.

Two R functions, $cox.zph()$ and $ggcoxzph()$, were used for the computation of GOF. The function $cox.zph()$ in the R survival package can test the relationship between the corresponding set of scaled Schoenfeld residuals and the failure time for each variable. The function $ggcoxzph()$ in the survminer package can produce graphs of the Schoenfeld residuals for each covariate against the time. R code for the check of Cox

PH assumption is listed in Code Snippet 4.4. The outputs regarding covariates in the fitted Cox model are obtained, see Figure 4.4 and Figure 4.5.

Figure 4.4 is the graphical output for risk factors: age (age), sex (sex), body mass index (bmi), and hypertension (hyp).

Figure 4.5 plots the graphical output for risk factors: systolic blood pressure (bps), cigarettes per day (cgrpd), pulse rate (pr), and the status of diabetes (dia).

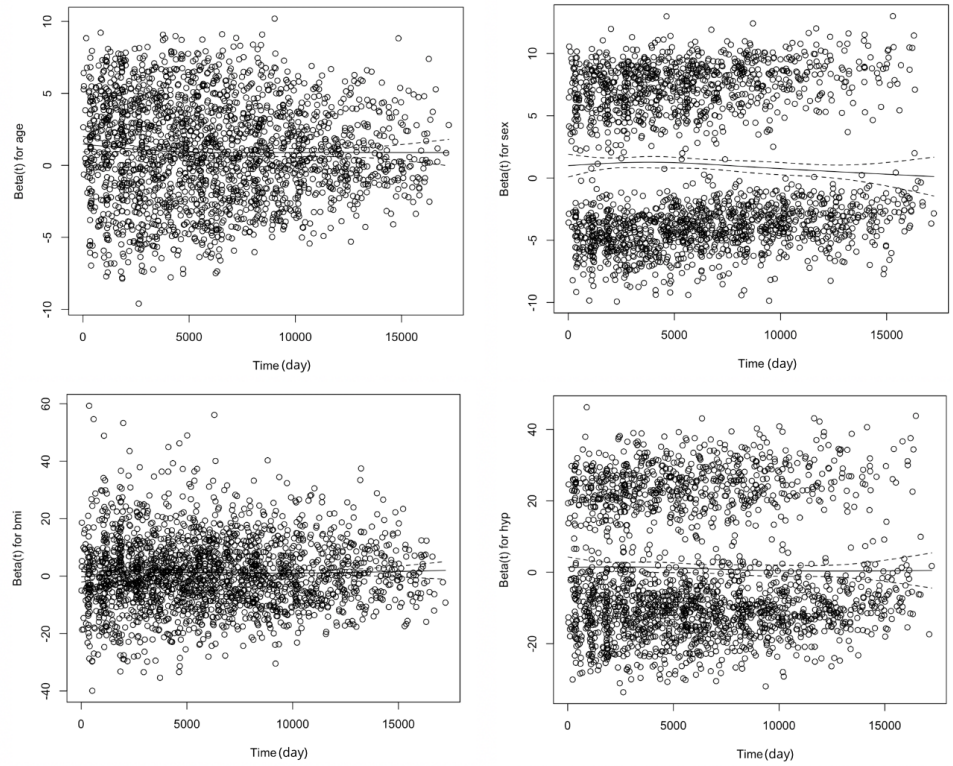


Figure 4.4: The Output of the PH Assumption Test: 1

For the graph of Schoenfeld test belonging to each covariate:

- The solid line in middle of the plot is a smoothing spline fit to the residuals.
- Two dotted lines separately displaying above and below the solid line represent a plus/minus error band for the fit.
- Values of y axes are estimated Schoenfeld residuals Beta over time.

According to Figure 4.4 and Figure 4.5, the PH assumption is supported by the

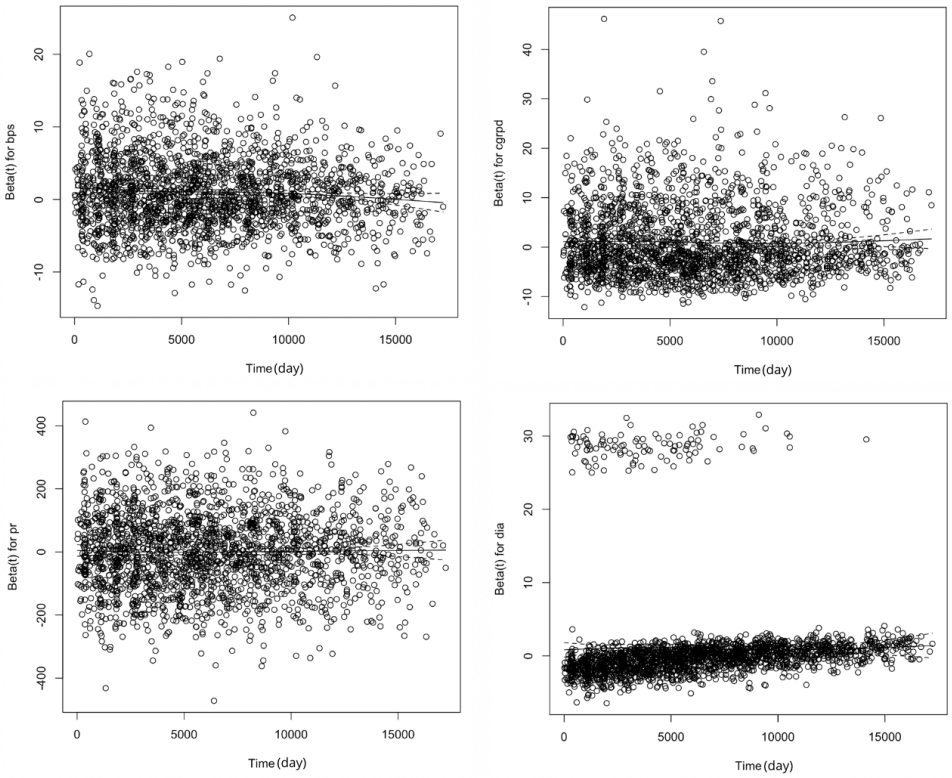


Figure 4.5: The Output of the PH Assumption Test: 2

non-obvious pattern between the Schoenfeld residuals and time. That means that if the estimated Beta coefficients systematically depart from a horizontal line, the assumption of PH will be violated. By inspecting the graphs depicted above, the assumption appears to be supported for all covariates, as there is not an obvious relationship between residuals. Time and the estimated 'Betas' for each covariate do not vary much over time. Taking factor "sex" as an example, it is a two-level factor, the two bands Betas do not obviously change over time.

The statistical test "p" value and "chisq" value are printed in Table 4.9. The "chisq" denotes the chi-square for each variable, i.e. the $\chi 2$ scale computed. A smaller "p" value has a larger "chisq" value. When using p-value to check the PH assumption for each variable, a large p-value, let's say > 0.1, indicates that the PH assumption is supported, whereas a very small p-value, say <0.05, suggests that the PH assumption is violated

Table 4.9: Statistical Output of the PH Assumption Test

Predictors	chisq	p
AGE	2.5862	0.1078
SEX	3.5131	0.06089
BMI	1.1499	0.28357
HYPERTENSION	1.0246	0.31144
BLOOD PRESSURE SYSTOLIC	0.0445	0.83299
CIGARETTES PER DAY	0.1577	0.6913
PULSE RATE	0.051	0.82132
DIABETES	0.0454	0.83136
GLOBAL	**27.0622**	**0.069**

(Kleinbaum & Klein, 2010).

According to results shown in Table 4.9, p-values of all variables are quite high, suggesting that all variables in the Cox-based risk model satisfy the PH assumption. The p-value for "sex" 0.06 is approximate to 0.05 but here we can assume that the PH assumption is reasonable for "sex". The result of GOF test referring p-values for all variables is consistent with the result of graphical check for the Cox PH assumption.

4.8 Summary

In this chapter, we developed a prognostic CVD prediction model using Cox regression analysis. Specifically, we first introduced the basic theory of the Cox regression model and then conducted a whole process of model development referring to the regression strategies, including the simulation of missing values using the multiple imputation method, variable selection using forward variable selection method, mutivariable Cox analysis and the development of the final model. After that, the baseline hazards at all covariates of mean values and zero were computed. Lastly, we did an evaluation of the Cox PH assumption test. Results of the assumption test held the PH assumption.

Chapter 5

Validation of the Cox-based Risk Prediction Model

5.1 Introduction

A prognostic fitted model cannot be applied to the practice unless it has been validated performing accurate prediction, i.e the GOF of the model, even though a data reduction (variable selection) method is used (Moons et al., 2009). Conducting a model validation is a method to make sure that a fitted diagnosis risk model was not over-fitted or otherwise was accurate. In other words, it is necessary to do a model validation to find out if predicted values from a fitted model can accurately predict future subjects that are not used to develop this model. Obtaining a high degree of certainty is particularly important when a model is developed for diagnosis or prognostic purpose.

The term "validation" has been constantly used in bio-statistics, machine learning, and artificial intelligence, etc. but seldom clarified (Feinstein, 1996). Altman and Royston (2000) thought that validating a prognostic model is the process of evaluating the performance of the model, while Katz (2011) pointed out that "validation" is the idea of proving that the inferences of establishing the model are true.

Further, we need to know the accepted standard to what extent the validation should be done. When fitting a model, it is to obtain the maximum probability of the occurrence of an outcome or event by analysing the values of the original data. However, the model might not perform as well with future new incoming data as with the original data. Successful validation is to verify the decrement in performance within a reasonable scale (Katz, 2011).

We applied two approaches to do the validation of our fitted model. We firstly used a classical statistical approach to get quantities of the model performance in internal data set (Exam 1), and then did an empirical validation by comparing the risk estimation using our model and the famous Framingham CVD prediction model (D'Agostino et al., 2008) for a series of selected individuals in the external data set (from Exam 8 to Exam 12 as shown in Table 3.2).

5.2 Statistical Validation

5.2.1 Methods of Statistical Validation

Within statistical validation, mainly two major modes of validation are commonly employed: internal and external (F. Harrell, 2013). Internal validation is considered the simplest validation method and therefore is widely used for assessing the performance of a fitted model on further samples. This method splits the original sample into two parts (Altman & Royston, 2000). The first proportion of the data is used to train the model (often called the training set) and the second proportion of the data is used for evaluating the model's ability to predict an event or outcome (called the test set).

The data-splitting can be done in various ways. It can be done randomly or in a non-random way. We could consider the distributions of the response and predictors in two sets of samples or not. Cross-validation, bootstrapping or other resampling

methods can be used in the procedures of data-splitting. Cross-validation has the ability to obtain an unbiased estimation of the predictive accuracy but tends to be imprecise, as what proportion of samples will be in the training and the test sets is randomly decided (Cox, 1975; Efron & Tibshirani, 1994). A tougher way to do the data splitting is in a non-random way. Compared with the cross-validation, bootstrapping is a better and more powerful approach to do this. Using bootstrapping, not only shrinkage factors can be estimated but also can be applied to the regression coefficients to avoid overoptimism (F. Harrell, 2013). In some cases, leave-one-out cross-validation (Schumacher, Holländer & Sauerbrei, 1997) or a chronological split (F. Harrell, 2013) can be used to get a prospective validation.

Compared with internal validation, external validation is more stringent. It evaluates a model on a future data set, sometimes collected from a population in a different place (Royston & Altman, 2013). Similar research design issues in the modelling process will be done, including sample selection, data measurements, etc. and need long time follow-up as well but external validation is a desirable and essential step before applying a model in clinical practice (Altman & Royston, 2000).

At this stage, limited to the time, we chose an internal validation method to validate our model. Bootstrapping will be used. A comparison of outcomes with these two methods will be described below.

5.2.2 Sample Size Considerations

Similar to the process of fitting models, the sample size and number of events should be considered for precise validation no matter which method will be used to validate our fitted model (Royston & Altman, 2013). The guiding principles in the model derivation data set also apply to the validation data. This means that substantial validation test samples are required.

For survival studies, only a few tens of patients is not large enough to validate a model. F. Harrell (2013) says that the outcome events in the validation sample set should not be less than 100. The research on the necessary sample size for validation is still ongoing (Jinks, 2012). It has been proven that bootstrap is an effective method for resampling the data set (Dunkler, Michiels & Schemper, 2007).

The sample size and number of events are summarised in section 3.3.1 indicating that we have 5079 samples and 3189 events in our data set. When applying the cross-validation and bootstrap methods into the validation process, the data size is large enough for us have a reasonable validation for our model.

5.2.3 Standards of Statistical Validation

When using statistical approaches for validation, two aspects should be verified: discrimination and calibration. A valid model should achieve satisfactory results of these two aspects in the validation sample set.

5.2.4 Discrimination of the Model

Discrimination, also called "separation", is the model's capability to discriminate outcomes of subjects. Predicting the incidence of high-risk patients should be higher than low-risk patients (Royston & Altman, 2013). The discrimination ability can be quantified in various statistical indexes such as Somers' D_{xy} (Somers, 1962), model χ^2 (Grambsch & O'Brien, 1991), receiver operating characteristic (ROC) area (Hanley & McNeil, 1982) or Spearman's ρ (Kendall, 1955), etc.

Among these statistics, the ROC area is the most commonly accepted measure of diagnostic discrimination (Hanley & McNeil, 1982). It is proven that the ROC

area is identical to the concordance index (c-index or c-statistics) (Austin & Steyerberg, 2017; Hanley & McNeil, 1982). The c-index computed based on the Wilcoxon–Mann–Whitney two-sample rank test (Whitehead, 1993) is the probability of a randomly selected patient who experienced an event, i.e. the concordance between predicted probability rate and response. The value of c-index can range from 0.5 to 1. A c-index of 0.5 indicates the result of the prediction is randomly obtained and a value of 1 suggests perfect prediction. Details of indication with different levels of this c-index value are explained in Table 5.1.

Table 5.1: Indications of the c-index with Different Values (Hanley & McNeil, 1982)

Value of c-index	**Indications**
= 0.5	A model with a random prediction
>0.7	A good model
>0.8	A strong model
= 1	A model with perfect prediction

Another statistical index used for assessing the model discrimination is Somers' D_{xy}, which measures the strength of the rank correlation between the actual observations and predicted probabilities (Somers, 1962). D_{xy} is also related to the c-index described above. Relations between D_{xy} and c-index are shown in Equation 5.1. When D_{xy} equals to 1, the c-index equals to 1, the model does perfect discrimination.

$$D_{xy} = 2 * (c - 0.5) \tag{5.1}$$

We use the R function $validate$ in rms package to assess the model's discrimination, the corresponding R code is listed below:

```
set.seed(1)
val <- validate(res.cph, method = "crossvalidation", u=10*365, B=150)
latex(val, file='') # print validation result into latex code
```

Code Snippet 5.1: R Code of Discrimination Assessment

Table 5.2: Output of Discrimination: Bootstrap

Index	Original Sample	Training Sample	Test Sample	Optimism	Corrected Index	n
D_{xy}	0.4171	0.4167	0.4149	0.0017	0.4154	150
Slope	1.0000	1.0000	0.9824	0.0176	0.9824	150
D	0.0344	0.0345	0.0339	0.0006	0.0338	150
U	−0.0001	−0.0001	0.0003	−0.0004	0.0003	150
Q	0.0346	0.0346	0.0336	0.0010	0.0336	150

The output of the discrimination code is shown in Table 5.2. According to the discrimination output, we have a value of Somers' D_{xy} 0.4149 for our model including these factors, equivalent to a ROC area of 0.70745 suggesting good discrimination.

Statistical indexes of predictive accuracy and the interpretations of these indexes are listed below:

- D_{xy} represents the Somers' rank correlation. It equals to $2(C - 0.5)$; C denotes c-index ("ROC Area" or the concordance).
- Slope is the calibration slope (predicted log odds versus true log odds). The slope is from (0, 1).
- D indicates the discrimination index. It equals to the likelihood ratio $\chi 2$ divided by the sample size.
- U is the index of unreliability indicating the extent that the logit calibration curve intercepts.
- Q is the accuracy score of logarithm. It is a scaled version of the log-likelihood.

5.2.5 Calibration of the Model

Calibration is the model's ability to estimate outcome without bias, which is another aspect for assessing the accuracy of a prognostic model. It indicates the reliability

of the model and reflects the predictive performance (Royston & Altman, 2013). A model under-predicting or over-predicting the event probability is miscalibrated but if well-calibrated, the event probability can be correctly assigned no matter what level of the predicted risk (F. Harrell, 2013).

A couple of approaches are available to assess the prediction accuracy of a model. One method is to compare the predicted and observed rates for individuals. This measure assesses the absolute prediction accuracy (Royston & Altman, 2013). A R function $calibrate$ is used to assess calibration of the model. Bootstrap was used as the internal resampling method. Graphical curves of calibration are shown in Figure 5.1. In the curve, errors are summarised by quantities mean absolute calibration error (0.009) and 0.9 quantiles calibration error (0.012).

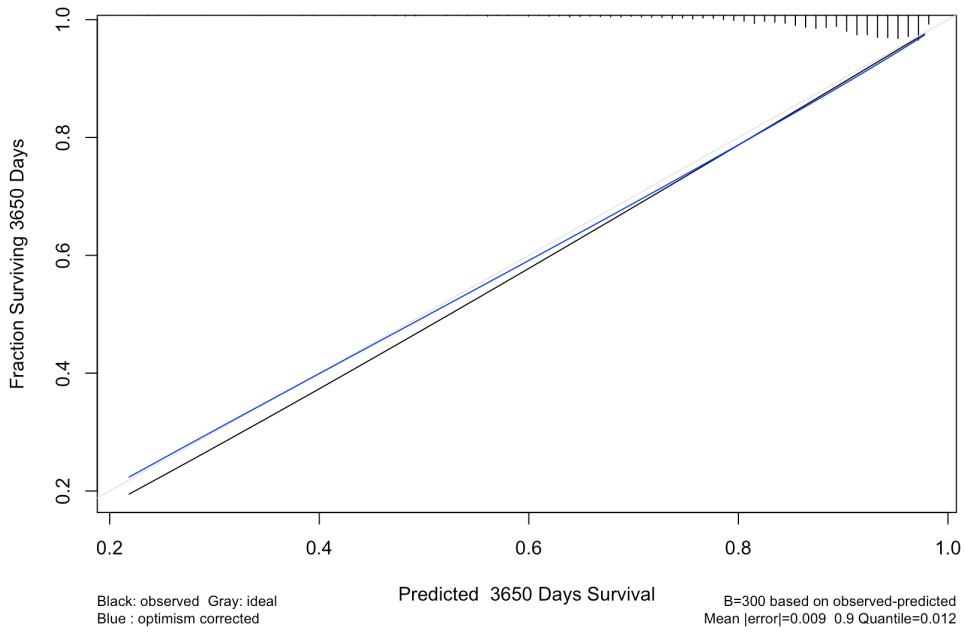


Figure 5.1: The Output of Calibration: Bootstrap

Figure 5.1 plots the calibration graph using bootstrap. The graph shows the evidence of overfitting the model where the model underestimates the low probabilities with a mean error 0.009 in the range of 0.2 to 0.6. Mean error is small and the c-statistic

derived in the section 5.2.4 is good (0.71). Combining the discrimination and the calibration results, the model is valid overall. The test results prove a goodness of the model fitting.

To further prove the effectiveness of the model, an empirical validation was done as follows.

5.3 Empirical Validation

We will compare the predictive ability of our risk model with one well-known CVD prediction tool, the Framingham prognostic model for general CVD (D'Agostino et al., 2008). We chose this model for comparison because it has been used for a long time, and has been validated several times in different populations (Hua et al., 2017; Hermansson & Kahan, 2018; Le, Marchant, Subtil, Boissel & Gueyffier, 2017). Its ability for estimating general CVDs as well as specific CVDs like stroke, heart attack, etc. has been widely recognised.

There are some differences on predictors used in our Cox-based model and the Framingham model. Specific predictors and their differences are listed in Table 5.3. Data regarding the predictors belonging to each model used for the empirical validation will be extracted and pre-processed using Python. The Python script can be accessed in Appendix D.

Next, we will implement the empirical validation in two perspectives:

- Horizontal comparison: comparing with the Framingham prognostic model using data collected from multiple samples at the same time-point.
- Longitudinal comparison: comparing with the Framingham prognostic model using data collected from specific samples at different time-points (fixed time intervals follow-up) and see the risk trend for an individual over time.

Table 5.3: Predictors in the Cox-based Model and the FHS Model

Predictors	Cox Model	Framingham Model
AGE	✓	✓
SEX	✓	✘
BMI	✓	✓
HYPERTENSION	✓	✘
TREATMENT OF HYPERTENSION	✘	✓
BLOOD PRESSURE SYSTOLIC	✓	✓
CIGARETTES PER DAY	✓	✘
SMOKING	✘	✓
PULSE RATE	✓	✘
DIABETES	✓	✓
Total	**8**	**6**

5.3.1 Horizontal Comparison

Samples used for the horizontal empirical validation are selected from the eighth exam data frame (Exam 8) from the Framingham Original Cohort Study. There are 2786 samples in this data frame where 1693 of them finally developed a CVD event. The event number is greater than the minimum number of effective validation (100 samples) so the sample size is large enough for a valid validation. Characteristics of samples used in the empirical validation are presented in Table 5.4.

Table 5.4: Samples and Events of the Horizontal Empirical Validation Data Set

Gender	Numbers	Events	Age Range
Male	1196	776	43-72
Female	1590	917	42-73
Total	2786	1693	42-73

As presented in section 3.3, 32 exam data frames were gathered in the Framingham heart study. The first exam (Exam 1) data has been used for model fitting. We chose Exam 8 data for an empirical validation because since then the data of a predictor (treatment of hypertension) in the Framingham model was collected. The Exam 8 data were collected from the same population in Exam 1 but at different time points. The

data in Exam 8 and Exam 1 are different. We think this horizontal comparison is an external validation within the same population.

Risks of developing CVD in 10 years for each sample using the Cox-based model and the Framingham model are estimated. For example, a subject with ID number 15018644, has a risk 12.57% using the Cox-based model and a risk 11.86% using the Framingham model. Values of predictors and risks for the subject 15018644 are shown in Table 5.5.

Table 5.5: Data Summary for Subject 15018644

PREDICTORS	VALUES	MEANINGS
AGE	44	YEARS
SEX	1	MALE
BMI	26.38689413	KG/M2
HYPERTENSION	0	NO
TREATMENT OF HYPERTENSION	0	NO
BLOOD PRESSURE SYSTOLIC	120	MM HG
CIGARETTES PER DAY	40	LAPSE
SMOKING	1	YES
PULSE RATE	82	PER MINUTE
DIABETES	0	NO
COX MODEL RISK	**12.57%**	
FHS MODEL RISK	**11.86%**	

Statistics of *min (lower whisker)*, *1st quartile (the lower hinge)*, *median*, *3rd quartile (the upper hinge)*, and *max (the extreme of the upper whisker)* of estimated risks for all samples are depicted in Figure 5.2. This box-whisker graph shows that risks estimated by our Cox-based model are higher than the risk calculated by the Framingham model but the error for five statistics (min, 1st Qu, median, mean, 3rd Qu., max) is within 0.02. For example, the median values of the FHS model and Cox-based model are correspondingly 0.1429475 and 0.1661985, and 0.1661985 is 0.023251 larger than 0.1429475.

To measure the accuracy of a risk model, we define a deviation z between the risk

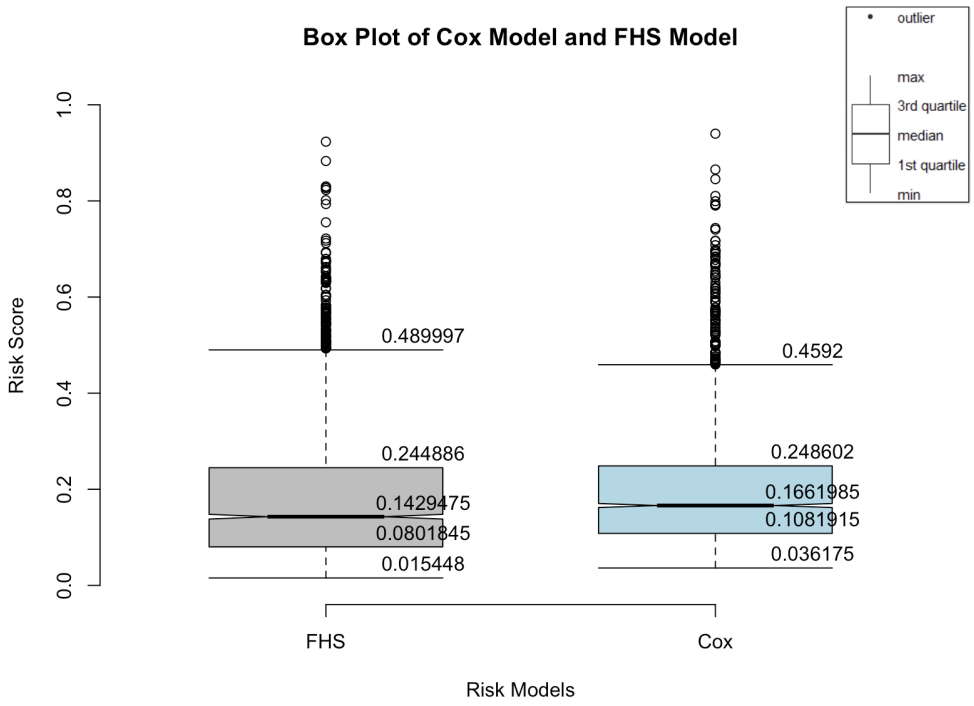


Figure 5.2: Horizontal Comparison between Cox-based Model and FHS Model

rate and the CVD occurrence, as shown below:

$$z = |CVD\ event - risk\ rate| \tag{5.2}$$

where $CVDevent$ is a boolean indicator to represent a CVD event. It could be 1 (with CVD) or 0 (without CVD); $riskrate$ is the risk estimate from a risk prediction model; the value z is the abs of $riskrate$ (the estimated risk score) minus the $CVDevent$ (0 or 1). When applying the FHS model and the Cox-based model to a subject, it is expected that a lower "z" value indicates a more accurate predictive ability.

The mean of z values for risk estimates from the FHS model and the Cox-based model are computed, as shown in Table 5.6. According to this table, the mean of z value for subjects who developed CVD using the Cox-based model (0.792) is lower than the FHS model (0.806), which indicates a much more accurate estimation. However, the mean of z value for subjects without CVD using the FHS model (0.162) is lower

Table 5.6: "z" Value Comparison between the FHS model and Cox-based Model

Samples	Num. of Samples	z (FHS Score)*	z (Cox Score)*
With CVD	1693	0.806	0.792
Without CVD	1093	0.162	0.172
Total	2786	0.484	0.482
*Notes: z = |cvd - risk score|, where cvd = 1 when sample has cvd; cvd = 0 when sample has not cvd			

than the Cox model (0.172), which indicates an overestimate of the Cox-based model. Overall, the predictive ability of these two models is consistent, as the total mean of z values for these two models are correspondingly 0.484 and 0.482. The risk scale of the Cox-based model is consistent with the Framingham model, which proves that our model is valid from another perspective.

5.3.2 Longitudinal Comparison

The longitudinal comparison focuses on the risk estimation for specific subjects but with different CVD scenarios. We selected four sex-specific subjects with or without CVD at the end of the Framingham Study. Characteristics of these four subjects are summarised in Table 5.7.

Table 5.7: Data Summary for Samples in the Longitudinal Validation

Samples	Gender	CVD	Diabetes
Sample 1	Male	✘	✘
Sample 2	Male	✓	✓
Sample 3	Female	✘	✘
Sample 4	Female	✓	✓

For each sample, data with fixed time intervals (approximately two years) from longitudinal time follow-up was extracted. Totally, five exams data (Exam 8, Exam 9,

Table 5.8: Exam Data for Sample 1: Male without CVD

Exams	age	bmi	bps	pr	cgrpd	trt	hyp	dia	smk
Exam 8	44	26.3868941	120	82	40	0	0	0	1
Exam 9	45	26.8266757	120	80	0	0	0	0	0
Exam 10	47	27.4676435	118	70	20	0	0	0	1
Exam 11	49	28.2222491	110	76	44	0	0	0	1
Exam 12	52	28.6750124	110	80	50	0	0	0	1

Table 5.9: Exam Data for Sample 2: Male with CVD and Diabetes

Exams	age	bmi	bps	pr	cgrpd	trt	hyp	dia	smk
Exam 8	45	27.7425797	132	83	20	0	0	0	1
Exam 9	47	26.2611798	124	80	20	0	0	0	1
Exam 10	49	27.6643519	130	78	20	0	1	0	1
Exam 11	51	27.1219136	130	90	20	0	1	0	1
Exam 12	53	24.8165509	122	82	20	0	0	1	1

Table 5.10: Exam Data for Sample 3: Female without CVD

Exams	age	bmi	bps	pr	cgrpd	trt	hyp	dia	smk
Exam 8	44	20.7763327	110	70	20	0	0	0	1
Exam 9	46	20.2654393	120	70	20	0	0	0	1
Exam 10	48	22.3120117	118	73	20	0	0	0	1
Exam 11	50	21.7971191	114	82	20	0	0	0	1
Exam 12	52	21.7971191	130	76	20	0	0	0	1

Table 5.11: Exam Data for Sample 4: Female with CVD and Diabetes

Exams	age	bmi	bps	pr	cgrpd	trt	hyp	dia	smk
Exam 8	46	21.7930436	130	65	3	0	1	0	1
Exam 9	48	21.9673879	170	75	16	0	1	0	1
Exam 10	50	22.494583	140	60	8	0	1	0	1
Exam 11	53	22.3174603	140	63	8	0	1	0	1
Exam 12	54	23.3801965	160	58	2	1	1	1	1

Exam 10, Exam 11, and Exam 12) were extracted for the longitudinal comparison. Data summaries for sample 1, sample 2, sample 3 and sample 4 are correspondingly listed in Table 5.8, Table 5.9, Table 5.10 and Table 5.11.

For each sample, risks of developing CVD in 10 years regarding five exams data were separately computed using our Cox-based model and the Framingham model. The

trend of risks over years is depicted with 5% error, as shown in Figure 5.3. From this graph, we can see that the trends of risks of these two models are consistent and risks for a specific sample increase over time, see the dotted trend lines in each graph. In addition, samples (both male and female) with diabetes that finally developed CVD have a higher risk rate than the ones who didn't develop the disease.

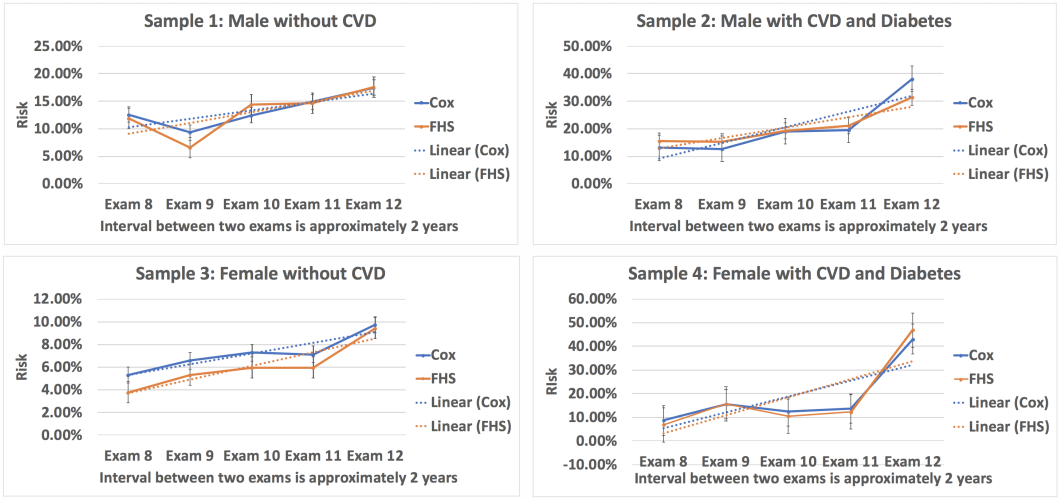


Figure 5.3: Longitudinal Validation

Overall, after conducting a horizontal and longitudinal empirical validation, the estimation ability of the Cox model is proved valid using external data sets. There is an error between the Cox-based model and the Framingham prognostic model but it is within the scale of 0.02.

5.4 Summary

In summary, we first implemented a statistical test for the GOF of our fitted model. The test results show that this model can complete a good discrimination and calibration achieving an area under cover for ROC equal to 0.71. We conducted an empirical validation by comparing the estimation ability of this model with the Framingham CVD model. The Framingham model has been clinically validated, so the empirical

validation can be regarded as a transformed clinical validation to some degree. A model validated statistically may not work well clinically but a model validated clinically may fail the statistical GOF test (Royston & Altman, 2013). Our model was both statistically validated and clinically validated and the results of these two kinds of validation indicate the goodness of our model.

Chapter 6

Findings and Discussion

6.1 Introduction

In this chapter, results and findings of this research are presented. After that, we discuss the Cox-based CVD risk prediction model we developed, and summarise the contribution of the research. Sections will be organised as follows:

- Section 6.2 has two parts. The first part is the presentation of our risk prediction model for general CVD, including key risk factors and the work-flow of computing the risk score. The second part is the survival estimation using nomograms and survival curves.
- Section 6.3 is a discussion of this research, including the indication of findings, comparison with other CVD risk prediction tools, and the main contributions.
- Section 6.4 is a short statement of this chapter.

6.2 Findings

6.2.1 Risk Factors in the Risk Model

Using the Cox regression analysis modelling method, a risk model including risk factors (as shown in Table 6.1) was developed. Statistics of "Min.", "1st Qu.", "Median", "Mean", "3rd Qu.", and "Max." are summarised. Variables included in this risk model are age, sex, BMI, hypertension, SBP, cigarettes per day, pulse rate, diabetes.

Table 6.1: Characteristics of Risk Factors Used in the Cox-based Risk Model

Predictors	Variables	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
AGE	age	28	37	44	44.15	51	74
SEX	sex	1	1	2	1.548	2	2
BMI	bmi	14.12	22.66	25.17	25.61	27.92	56.68
HYPERTENSION	hyp	0	0	0	0.147	0	1
BLOOD PRESSURE SYSTOLIC	bps	84	122	136	138.6	150	270
CIGARETTES PER DAY	cgrpd	0	5	20	16.26	20	60
PULSE RATE	pr	37	67	75	75.61	83	170
DIABETES	dia	0	0	0	0.0197	0	1

6.2.2 General CVD Risk Prediction Model

The regression coefficients, HRs (Hazard Ratios), lower .95, and upper .95 confidence interval for each predictor in our general CVD risk model are presented in Table 6.2. For continuous variables, the regression coefficients and hazard ratios are for their natural logarithms.

As the Cox model survival function has the form:

$$S(t) = [S_0(t)]^{exp(\sum_{i=1}^{k} \beta_i X_i)} \tag{6.1}$$

the Cox hazard model can be written as a general formula:

$$\hat{H(t)} = 1 - S_0(t)^{exp(\sum_{i=1}^{k} \beta_i X_i - \sum_{i=1}^{k} \beta_i \bar{X}_i)} \tag{6.2}$$

Table 6.2: Regression Coefficients and Hazard Ratios in the Cox-based Risk Model

Predictors	Variables	coef*	Hazard Ratio	lower .95	upper .95
AGE	log of age	2.083643	8.033686	6.4082	10.0716
SEX	sex	-0.469719	0.625178	0.5787	0.6754
BMI	log of bmi	0.608864	1.838342	1.4368	2.3521
HYPERTENSION	hyp	0.241461	1.273108	1.1342	1.429
BLOOD PRESSURE SYSTOLIC	log of bps	1.682571	5.37937	3.7938	7.6277
CIGARETTES PER DAY	cgrpd	0.009669	1.009716	1.0065	1.013
PULSE RATE	log of pr	-0.30209	0.739271	0.5879	0.9297
DIABETES	dia	1.087501	2.96685	2.3244	3.7869
*Estimated regression coefficient					

where $\hat{H(t)}$ is the CVD risk estimated for an individual; $S_0(t)$ is baseline survival rate at time t, here t = 10 (years), see Table 4.8; β_i is the regression coefficient, see the column "coef*" in Table 6.2; X_i is the value of the ith variable, if continuous it is the log-transformed value); $\bar{X}_i$ is the mean value of the set of the subject i, and k denotes the number of predictive variables.

6.2.3 10-year Risk Score Computation

Using the CVD risk function (see Equation 6.5) and regression coefficients (see Table 6.2), we can easily compute the probability of developing any type of CVD for an individual.

Here, we take the sample with the ID number 15018644 (once used as an example in horizontal validation in Section 5.3.1) as a specific case to illustrate the process of risk score calculation. As shown in Table 5.5, this sample is a 44-year-old man not having diabetes and hypertension. He has a systolic blood pressure of 120 mmHg, pulse rate of 82 per minute, BMI of 26.38689413 kg/m_2, and is a current smoker smoking 40 lapses per day.

The CVD risk using the Cox-based risk model is calculated as follows:

$$\begin{aligned}\sum_{i=1}^{k} \beta_i X_i = 2.083643 * log(44) - 0.469719 * 1 + 0.608864 * log(26.38689413) \\ +0.241461 * 0 + 1.682571 * log(120) - 0.302090 * log(82) \\ +0.009669 * 40 + 1.087501 * 0 \\ = 16.518741\end{aligned} \tag{6.3}$$

$$\begin{aligned}\sum_{i=1}^{k} \beta_i \bar{X}_i = 2.083643 * 3.768 - 0.469719 * 1.548 + 0.608864 * 3.230 \\ +0.241461 * 0.1469 + 1.682571 * 4.913 - 0.302090 * 4.311 \\ +0.009669 * 13.96 + 1.087501 * 0.02001 \\ = 16.247045\end{aligned} \tag{6.4}$$

$$\begin{aligned}\hat{H(10)} &= 1 - S_0(10)^{exp(\sum_{i=1}^{k} \beta_i X_i - \sum_{i=1}^{k} \beta_i \bar{X}_i)} \\ &= 1 - 0.9027267^{exp(16.518741 - 16.247045)} \\ &= 0.125658 \approx 12.57\%\end{aligned} \tag{6.5}$$

The Cox-based risk model gives a 10-year estimate of 12.57% for the sample with the ID number 15018644.

6.2.4 Survival Estimation

Apart from estimating the CVD hazard rate using the general formula as shown in Equation 6.5, graphs for the survival estimation can be plotted. Below we will introduce two forms of survival estimation, namely, the nomogram and survival curves.

6.2.4.1 Nomogram

A nomogram is a two-dimensional diagram to represent a mathematical function involving several predictors (Kattan, 2003b). It is a simple graphical computing device to approximately estimate the probability of a particular event based on statistical regression methods such as the Cox regression model for survival analysis (Kattan, 2003a).

Nomograms have become very popular among clinicians, primarily because they are designed to provide individualised estimates of the probability of a specific event such as the occurrence of CVD. They can help clinicians or individuals have a prognosis of the event based on the characteristics of the individual patients.

According to the step-by-step guidance for building a nomogram from a regression fit (Iasonos, Schrag, Raj & Panageas, 2008), a nomogram from the risk prediction model we fitted in Chapter 4 was created. This nomogram implements the estimation of individual survivals in 10 years.

Key steps to create the nomogram are described as follow:

- Step 1: store the predicted survival from the fitted model
- Step 2: calculate the survival in 10 years
- Step 3: scale the survival probability
- Step 4: create the nomogram

The corresponding R code for building the nomogram is listed below:

```
# Step 1
surv  <- Survival(res.cph.exam1)

# Step 2
surv10 <- function(x) surv(10*365,lp=x)

```

```
#Step 3
scale    <- c(.05,.1,.2,.3,.4,.5,.6,.7,.8,.9,.95)

#Step 4
nomogram <- nomogram(res.cph.exam1,
                fun=list(surv10),
                funlabel=c('10-year Survival'),
                fun.at=list(scale, scale, c(.5,1:6)))

plot(nomogram, xfrac=.65, lmgp=.35)
```

Code Snippet 6.1: R Code for Building the Nomogram

Figure 6.1 is the nomogram after running the R code listed in Code Snippet 6.1. Briefly, this diagram creates a simple graphical representation of our CVD risk prediction model. Each predictor has a set of n scales and there is a mapping between each scale and the "Points" scale. The bottom is the 10-year survival estimates. By accumulating the total points corresponding to the specific configuration of covariates for a patient, a clinician can then manually obtain the predicted value of the event for that patient.

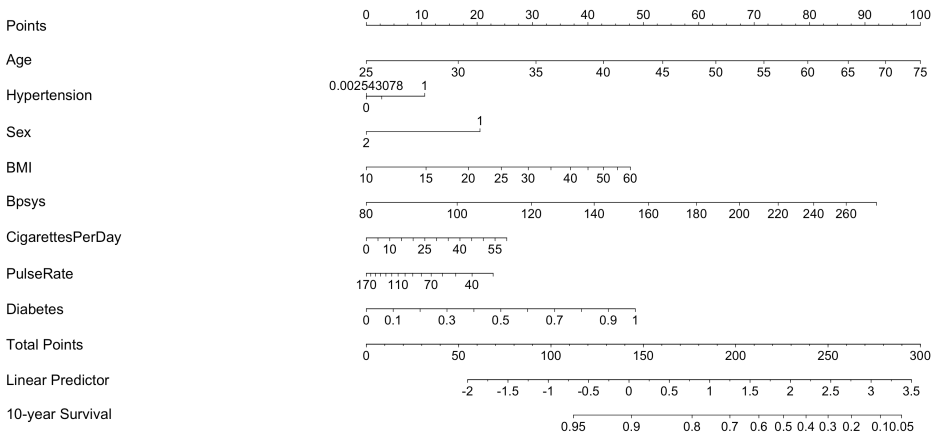


Figure 6.1: Nomogram for Predicting Overall Survival in 10 Years

Steps to read the nomogram shown in Figure 6.1,

- First, draw a vertical line across a predictor line to the top axis labelled "Points"; each tick marker on the line indicates the point value of a predictor.

- Second, sum points belonging to all predictors and find the corresponding value on the bottom axis labelled "Total Points".
- Third, draw a vertical line from the tick marker points on the "Total Points" down to the axis showing "10-year survival" and read the 10-year survival rate.

We still take the sample with ID number 15018644 as an example to calculate his 10-year survival using the nomogram in Figure 6.1. Following the steps of reading a nomogram, the corresponding points to each value of each predictor are mapped as shown in Table 5.5. For example, a systolic blood pressure of 120 mmHg has 30 points. After obtaining the points for each predictor the total points is summed as 155 points, as shown in Table 6.3. By mapping this total to the survival estimation scales, this man has a 87.5% probability of surviving without CVD in 10 years. Correspondingly, the hazard rate in 10 years is 12.5% (approximately equivalent to estimate rate calculated by the general Formula 6.5, which is 12.57%).

Table 6.3: Case: Results of Reading the Nomogram

Predictors	Values	Points
AGE	44	52.5
SEX	1	20
BMI	26.386894	25
HYPERTENSION	0	0
BLOOD PRESSURE SYSTOLIC	120	30
CIGARETTES PER DAY	40	17.5
PULSE RATE	82	10
DIABETES	0	0
Total Points	-	155
10-year survival (%)	87.50%	
10-year hazard (%)	12.50%	

A nomogram's performance is usually evaluated in terms of the discrimination and calibration (Kattan, 2003a), which we have done in Chapter 5. From the example we computed the survival probability using the nomogram we have built, the performance

of this diagram shows that it has high accuracy for predicting the survival probability or hazard rate for CVDs. This ensures that the nomogram would perform well when it is used in future subjects. These user-friendly nomograms facilitate the clinicians and patients to generate risk estimates for decision making.

6.2.4.2 Individual Survival Curves

A survival curve is a diagram statistically depicting the survival estimate of an individual or a group of patients in a given length of time. A percentage of surviving versus time is shown in the survival graph. The survival rate change over time can be seen intuitively and helps us to understand how the trend differs among different individuals. The characteristics of survival curves are important for clinicians to think about prognosis of some disease as well as treatment choices.

Here is an example to plot individualised survival probabilities using R function $survfit$ in $survival$ package. Sample data is still from the sample with the ID number 15018644 (Table 5.5). R code for extracting the survival curve for this person is listed below:

```
# data frame for the sample 15018644
new.data <- data.frame(Age=44,Hypertension=0,Sex=1,BMI=26.38689413,Bpsys=120,CigarettesPerDay=40,
           PulseRate=82,Diabetes=0)
# plot the survival curve
plot(survfit(res.cph.exam1, newdata=new.data),
     xscale=365.25, xlab = "Years", ylab="Survival")
```

Code Snippet 6.2: R Code for Extracting the Survival Curve

Figure 6.2 is the survival curve for the individual shown in Table 5.5. This plot shows the survival curves of specific values of the risk factors (age, sex, hypertension, BMI, SBP, cigarettes per day, pulse rate, and the status of diabetes) belonging to that individual. In Figure 6.2, the Y axis shows the actual percentage of surviving. The value runs from 1 at the top to 0 at the bottom, indicating approximately 100% survival

to 0% survival at the bottom. The X axis gives the time (years) from the start of the observation to the end of the experiment. The two dotted lines upper and lower the solid line are the mean error curves.

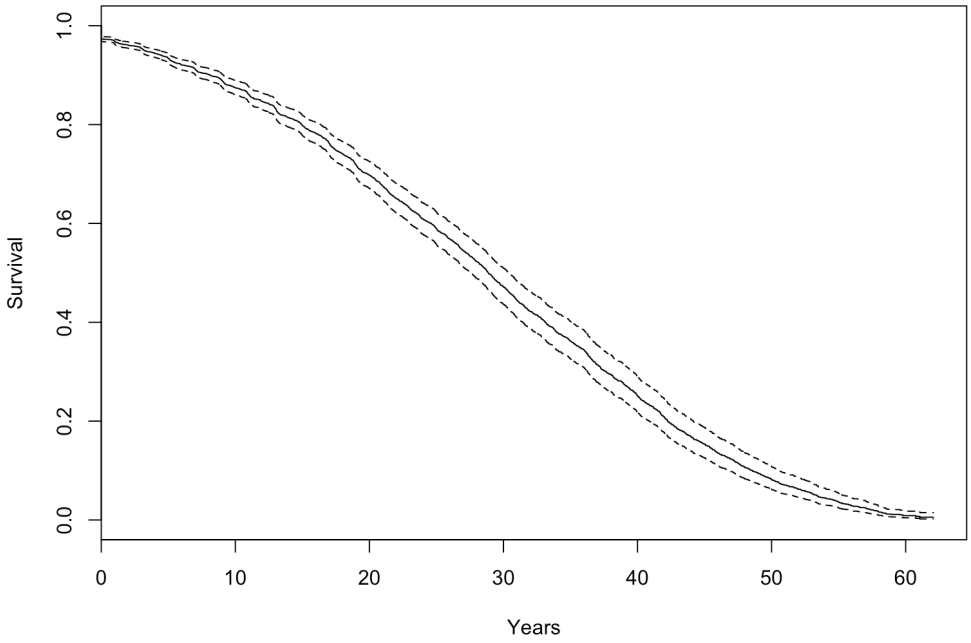


Figure 6.2: Individual Survival Curve for the Sample ID 15018644

It is easy to read the survival percentage by mapping the X axis value to the vertical point proportion of surviving. For example, if a clinician would like to know the survival probability at particular time of 10 years, he can easily obtain the percentage, approximately 0.875 (87.5%). This curve starts out with approximate 100% and descends over time but of course it can never increase.

6.3 Discussion

It is widely known that CVD has become one of the major public health problems globally (Lopez, Mathers, Ezzati, Jamison & Murray, 2006) and in New Zealand (Hay,

2004), and continues producing immense burdens for the health care and economic system in the community. Prior researchers have noted the importance of identifying associated risk predictors and the early detection and intervention of CVDs (Hubert, Feinleib, McNamara & Castelli, 1983; Cupples, 1987; Weiner et al., 2004; Böhm et al., 2010; Odden et al., 2014), and then reducing the risk of developing the diseases prior to the occurrence of the disease. Consequently, CVD risk prediction tools based on a single variable or multiple variables have been devised to yield estimates of the CVD risk (Wilson et al., 1998; Conroy et al., 2003; Hippisley-Cox et al., 2007; D'Agostino et al., 2008; De Ruijter et al., 2009; Pencina et al., 2009; Bannink, Wells, Broad, Riddell & Jackson, 2006).

Motivated by the objective of early detection and CVD risk estimation, the present research was designed to identify novel CVD risk factors, determine the effect of these factors, and then develop a risk prediction model based on the identified factors. Although risk factors could vary from one specific CVD component to another, there is sufficient evidence that different types of CVD have commonalities of risk factors. With respect to the aim of this research, we derived and validated a new 10-year risk model for general CVDs based on time follow-up data rigorously measured by the Framingham Heart Study researchers. It is expected this model can help clinicians and individuals to prevent the occurrence of CVD.

This investigation extends the number of risk factors on the basis of the previous general CVD risk formulations, incorporating heart rate to estimate absolute CVD risk. The approach used in this research is based on advanced statistical techniques that allow reducing the bias in the assessment of true CVD risk. The whole process of data analysis strictly follows the guideline of regression modelling strategies and survival analysis (Kleinbaum & Klein, 2010; F. Harrell, 2013). Our general CVD risk model shows a good discrimination and calibration (obtaining a c-statistic as 0.71).

We use continuous variables (age, BMI, SBP, pulse rate) to generate the model

that performs better than models developed using categorical variables. Compared with simpler approaches that try to make inferences of 10-year risk models such as the logistic-regression-based model developed by Kannel, McGee and Gordon (1976), and the CKD risk model using Kaplan-Meier and the log-rank test (Weiner et al., 2004), our risk model is more adequate and will not cause serious errors of underestimation or overestimation. Moreover, this model was developed based on a larger number of samples and events, suggesting a valid estimation of the true risk.

6.3.1 Comparison with Other CVD Risk Prediction Tools

The old version Framingham general CVD risk function formulated by Kannel et al. (1976) is effective for identifying persons at high risk of CVD but it was based on a limited number of risk factors (serum cholesterol, SBP, smoking history, electrocardiogram, and glucose intolerance). The Framingham investigators devised two new sex-specified general CVD risk functions recently (D'Agostino et al., 2008). One laboratory-test-based formula included HDL cholesterol and the other office-based one include BMI in the risk function. The QRISK study investigators incorporated family history as a novel risk factor on the basis of the Framingham general formulas (Hippisley-Cox et al., 2007). Although researchers have published risk scores (Kannel et al., 1976; D'Agostino et al., 2008; Hippisley-Cox et al., 2007) for predicting general CVDs, these functions did not include heart rate in the risk model.

Several risk models formulated by using machine learning or data mining techniques have incorporated heart rate as a risk factor but tools that can predict CVD absolute risk are fewer. For example, the prediction tool proposed by H. Kim et al. (2016) focuses on the classification of CVD event by employing ANN and the Bayesian classifier based on heart rate variability. The diagnosis CVD model developed by Vaanathi (2017) categorises the CVD risk as different levels but an absolute risk score cannot be obtained.

Even though the supportive tool developed by Unnikrishnan et al. (2016) will give an estimate of a risk score but the user can not know how many years the score is targeting.

In addition, there are some equations focusing on specific CVD outcomes. The Europe SCORE project offers equations to estimate a fatal cardiovascular event. There are risk estimation tools for coronary heart disease (Kannel et al., 1979; Wilson et al., 1998; Lloyd-Jones et al., 2004), for general heart disease (McGill et al., 2008), for stroke (Yu et al., 2017; Parmar et al., 2015), etc.. Compared with disease-specific models to estimate the risk of developing specific CVD outcomes, the present study generated a general CVD risk model that can predict a global CVD risk as well as the risk of developing individual components.

Moreover, compared with the laboratory-based algorithms, the present research proposed a simple way to estimate 10-year CVD risk based on office-measured risk factors, there being no need to do clinical or laboratory tests such as the HDL measurement or blood test. An individual can assess his or her CVD risk during an office visit or his own monitoring of the combination of risk factors in the risk model, either manually or using some device like a wearable sensor.

6.3.2 Contributions

The main contribution of this study is the risk prediction model for early detection of CVD, which is very encouraging. Findings of this research are significant in at least four major respects.

- Novel risk factor heart rate and conventional risk factors were identified as significant for the development of CVD.
- A simple office-based CVD risk prediction model aiming at general CVD was developed. This risk model does not require a laboratory test. Practitioners can use it during an office visit.

- Absolute risk scores in 10 years of CVD can be estimated.
- Multiple forms of the CVD risk estimation, namely the risk equation, nomogram, and survival curves, were provided.

Those findings may have significant impacts on different stakeholders in the following ways.

- To clinicians: this risk model would be an effective prognosis tool for the primary prevention of CVD. It helps them estimate an accurate probability of developing CVD in the next few years and provides a corresponding recommendation of life style intervention or clinical treatments.
- To individuals: this risk model would be a convenient home care tool. They can use this tool to self-monitor their risk changes at home when they are doing their improvements or treatment.
- To industries: this risk model can be applied to industries such as health care applications or wearable monitoring devices for office-based CVD prevention practice.

6.3.3 Implications

This finding has important implications for developing a wearable monitoring system for integrating the CVD risk prediction model into the primary health care practice. This would be a revolution in health care management and spending. However, with the small event size of diabetes, caution must be applied to the practice of this risk model. Even though we have used the multiple imputation method to impute the missing values for diabetes, the original feature of data imbalance, which decides that the imputed data frame for the "diabetes" might still have a data in-balance there. Advanced imputation

methods need to considered in the future for avoiding unexpected outcomes caused by the diabetes data imbalance.

As this risk profile aims to provide an office-based prediction tool for convenience, we did not include risk factors based on a clinical test such as total cholesterol and HDL cholesterol but they have a strong effect on CVD event. We have provided a valid framework for creating a risk model using the Cox regression model, but future work should take into account risk factors not included in our model at this moment. Thus, expanding more predictors into the risk model is an important issue for future research.

6.4 Summary

In summary, this section demonstrated the proposed general CVD risk prediction model in different ways. Both mathematical and graphical presentations of this proposed model were provided. For an individual sample, the estimated risks are consistent by using the risk equation, the nomogram, and the survival curves. Moreover, a discussion about the findings of this research indicates that the general CVD risk model can be an effective tool for the prevention of CVDs but the implications of the limitations of this research should be considered in future studies.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

In this research, we developed a risk prediction model based on office-based multivariable predictors. This risk model can be used for assessing the probability rate of developing general CVDs and specific CVD components such as the stroke, heart failure, and diabetes. A full process of developing a risk prediction model was demonstrated using the Cox regression method. Both statistical validation and empirical validation were conducted, and a satisfying discrimination and calibration ability (ROC being 0.71) were obtained. The risk score estimated from the prediction model we derived is considered to be a valid tool for practitioners to quantify CVD risk and help them identify high risk individuals for further preventive primary health care.

The findings of this research suggest that heart rate is a novel independent risk factor affecting the occurrence of CVD. Incorporating this predictor into the risk estimation model on the basis of traditional risk factors expands the predictive ability of past existing CVD risk equations. To our knowledge, this is the first algorithm that incorporates heart rate and traditional risk factors using a statistical method. Moreover, this office-based risk factor model (as shown in Chapter 6) does not require a clinical

test, so is more easily operated during an office visit. Both clinicians and patients can use it simply and easily to assess CVD event rates at any time and any place.

7.2 Limitations

Limitations of our study need to be acknowledged.

- Only a subset of the Framingham Heart Study data was selected for our model fitting. Among the data set, the size of diabetes events is small. This data imbalance of diabetes events might cause bias (unexpected outcome) when applying the proposed risk model to special diabetes populations.
- We obtained a result of discrimination and calibration as 0.71, which proves a good model but not a perfect model. Further research should focus on the improvement of the model's ability of discrimination and calibration.
- The validity of the proposed Cox-based risk model for general CVDs should be evaluated in different ethnic populations. We did an internal validation and an external empirical validation but within the same population. Stringent external validations should be conducted in future studies to make sure this model can be transported to different ethnicities.

7.3 Future Work

To avoid the bias caused by the limited size of diabetes events, future research should expand the size of data for data analysis. More data frames need to be included and pre-processed. Possible methods are listed below:

- We have three cohorts study data downloaded from the Framingham Heart Study

organisation. We can extract all data to expand the sample size for an adequate event size of diabetes.

- Access and request other available longitudinal data sets, such as the KCIS programme data set (Chiu et al., 2006), the PREDICT cohort study data set (Wells et al., 2017), or the QRISK cohort study data set (Hippisley-Cox et al., 2007).

Furthermore, future work should combine these presented risk prediction equations with real-time health care platforms. Real-time data frames are collected from a wearable sensor. Users can monitor their real-time CVD risk trend as well as survival curve changes. For achieving this objective, three components need to be set up: a wearable sensor that can accurately collect risk factors included in this risk model, a secure off-site server for storing and processing the real-time data, and a platform that can be run in a wearable device such as a smart phone, tablet, or smart watch.

To improve the validity of the model, efforts can be made in the data process stage if we continue to use statistical regression methods to fit the risk estimation model. In parametric or semi-parametric models, restricted cubic spline function (Devlin, Weeks et al., 1986) can be applied to continuous variables. Data will be normalised for reducing the degree of non-linearity of the predictor. Or else, advanced data mining and machine learning techniques can be applied to data analysis. Although these techniques have the disadvantages discussed in Section 1.2, they have the ability to increase the validity of discrimination and calibration of a risk model (Unnikrishnan et al., 2016).

In addition, there are marked differences between different ethnic groups. Risk models developed in one population might underestimate or overestimate the risk in another population. For instance, the Framingham score will overestimate the risk of CHD for the population in people of China (Zhang, Attia, D'Este, Yu & Wu, 2005). The accuracy and transportability of the generated CVD risk model should be externally

validated in different ethnic populations. For applying this CVD risk prediction model on New Zealanders, our next step could be requesting a cohort study data set (called 'PREDICT') from the New Zealand population (Wells et al., 2017), and then conducting an external validation on the basis of the PREDICT data set.

References

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, *16*(1), 125–127.

Altman, D. G. & Royston, P. (2000). What do we mean by validating a prognostic model? *Statistics in medicine*, *19*(4), 453–473.

Anderson, K. M., Odell, P. M., Wilson, P. W. & Kannel, W. B. (1991). Cardiovascular disease risk profiles. *American heart journal*, *121*(1), 293–298.

Assmann, G., Cullen, P. & Schulte, H. (2002). Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular münster (procam) study. *Circulation*, *105*(3), 310–315.

Austin, P. C. & Steyerberg, E. W. (2017). Events per variable (epv) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. *Statistical methods in medical research*, *26*(2), 796–808.

Bannink, L., Wells, L., Broad, J., Riddell, T. & Jackson, R. (2006). Web-based assessment of cardiovascular disease risk in routine primary care practice in New Zealand: the first 18,000 patients (PREDICT CVD-1).

Berry, M. J. & Linoff, G. (1997). *Data mining techniques: for marketing, sales, and customer support*. John Wiley & Sons, Inc.

Bhatla, N. & Jyoti, K. (2012). An analysis of heart disease prediction using different data mining techniques. *International Journal of Engineering*, *1*(8), 1–4.

Board, J. B. S. (2014). Joint British Societies' consensus recommendations for the prevention of cardiovascular disease (JBS3). *Heart*, *100*(Suppl 2), ii1–ii67.

Böhm, M., Swedberg, K., Komajda, M., Borer, J. S., Ford, I., Dubost-Brama, A., ... others (2010). Heart rate as a risk factor in chronic heart failure (shift): the association between heart rate and outcomes in a randomised placebo-controlled trial. *The Lancet*, *376*(9744), 886–894.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, 89–99.

Burke, L. E., Ma, J., Azar, K. M., Bennett, G. G., Peterson, E. D., Zheng, Y., ... others (2015). Current science on consumer use of mobile health for cardiovascular disease prevention. *Circulation*, CIR–0000000000000232.

Butler, J., Kalogeropoulos, A., Georgiopoulou, V., Belue, R., Rodondi, N., Garcia, M., ... Kritchevsky, S. B. (2008). Incident heart failure prediction in the elderlyclinical perspective. *Circulation: Heart Failure*, *1*(2), 125–133. doi: 10.1161/CIRCHEARTFAILURE.108.768457

Cannon, A. (2012). *Reliability data banks*. Springer Science & Business Media.

Cardiovascular Disease Risk Assessment Steering Group and others. (2017). *New Zealand primary care handbook 2012. wellington: Ministry of health; 2013.*

Carter, M. C., Burley, V. J., Nykjaer, C. & Cade, J. E. (2013). Adherence to a smartphone application for weight loss compared to website and paper diary: pilot randomized controlled trial. *Journal of medical Internet research*, *15*(4).

Centers for Disease Control and Prevention and others. (2009). *Blood glucose daily self-monitoring: rate of daily self-monitoring among adults with diabetes aged 18 years and older, 1997–2006.*

CG181, NICE. (2014). *Cardiovascular disease: risk assessment and reduction, including lipid modification.* July.

Chan, W. C., Wright, C., Riddell, T., Wells, S., Kerr, A. J., Gala, G. & Jackson, R. (2008). Ethnic and socioeconomic disparities in the prevalence of cardiovascular disease in New Zealand. *The New Zealand Medical Journal (Online)*, *121*(1285).

Chen, T., Chiu, Y., Luh, D., Yen, M., Wu, H., Chen, L., ... others (n.d.). Taiwan community-based integrated screening group (2004) community-based multiple screening model: design, implementation, and analysis of 42,387 participants. *Cancer*, *100*(8), 1734–43.

Chiu, Y.-H., Chen, L.-S., Chan, C.-C., Liou, D.-M., Wu, S.-C., Kuo, H.-S., ... Chen, T. H.-H. (2006). Health information system for community-based multiple screening in Keelung, Taiwan (Keelung community-based integrated screening no. 3). *International Journal of Medical Informatics*, *75*, 369 - 383.

Choi, S. & Jiang, Z. (2006). A novel wearable sensor device with conductive fabric and pvdf film for monitoring cardiorespiratory signals. *Sensors and Actuators A: Physical*, *128*(2), 317–326.

Chong, I.-G. & Jun, C.-H. (2005). Performance of some variable selection methods when multicollinearity is present. *Chemometrics and intelligent laboratory systems*, *78*(1-2), 103–112.

Collaboration, A. P. C. S. et al. (2004). Body mass index and cardiovascular disease in the Asia-Pacific region: an overview of 33 cohorts involving 310 000 participants. *International journal of epidemiology*, *33*(4), 751–758.

Conroy, R., Pyörälä, K., Fitzgerald, A. e., Sans, S., Menotti, A., De Backer, G., ... others (2003). Estimation of ten-year risk of fatal cardiovascular disease in europe: the score project. *European heart journal*, *24*(11), 987–1003.

Cox, D. R. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika*, *62*(2), 441–444.

Cox, D. R. (1992). Regression models and life-tables. In *Breakthroughs in statistics* (pp. 527–541). Springer.

Cox, D. R. & Oakes, D. (1984). *Analysis of survival data* (Vol. 21). CRC Press.

Creswell, J. W. (2014). *Research design : qualitative, quantitative, and mixed methods approaches.* Thousand Oaks : SAGE Publications, [2014].

Cui, J. (2009). Overview of risk prediction models in cardiovascular disease research. *Annals of epidemiology*, *19*(10), 711–717.

Cupples, L. (1987). Some risk factors related to the annual incidence of cardiovascular disease and death using pooled repeated biennial measurements. *Framingham Heart Study*.

D'Agostino Sr, R. B., Grundy, S., Sullivan, L. M., Wilson, P., Group, C. R. P. et al. (2001). Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. *Jama*, *286*(2), 180–187.

Damen, J. A., Hooft, L., Schuit, E., Debray, T. P., Collins, G. S., Tzoulaki, I., ... others (2016). Prediction models for cardiovascular disease risk in the general population: systematic review. *Bmj*, *353*, i2416.

Dawber, T. R., Kannel, W. B. & Lyell, L. P. (1963). An approach to longitudinal studies in a community: the Framingham study. *Annals of the New York Academy of sciences*, *107*(1), 539–556.

Denzin, N. K. & Lincoln, Y. S. (2011). *The sage handbook of qualitative research*. Sage.

De Ruijter, W., Westendorp, R. G., Assendelft, W. J., den Elzen, W. P., de Craen, A. J., le Cessie, S. & Gussekloo, J. (2009). Use of Framingham risk score and new biomarkers to predict cardiovascular mortality in older people: population based observational cohort study. *Bmj*, *338*, a3083.

Despa, S. (2010). What is survival analysis. *Cornell University, Cornell Statistical Consulting Unit, Newsletter*.

Devlin, T., Weeks, B. et al. (1986). Spline functions for logistic regression modeling. In *Proceedings of the 11th annual sas users group international conference* (Vol. 646, p. 51).

Dunkler, D., Michiels, S. & Schemper, M. (2007). Gene expression profiling: does it add predictive accuracy to clinical characteristics in cancer prognosis? *European Journal of Cancer*, *43*(4), 745–751.

D'Agostino, R. B., Vasan, R. S., Pencina, M. J., Wolf, P. A., Cobain, M., Massaro, J. M. & Kannel, W. B. (2008). General cardiovascular risk profile for use in primary care. *Circulation*, *117*(6), 743–753.

Efron, B. (1977). The efficiency of cox's likelihood function for censored data. *Journal of the American statistical Association*, *72*(359), 557–565.

Efron, B. & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.

Eng, T. R. (2001). The ehealth landscape: a terrain map of emerging information and communication technologies in health and health care.

Etemadi, M., Inan, O. T., Heller, J. A., Hersek, S., Klein, L. & Roy, S. (2016). A wearable patch to enable long-term monitoring of environmental, activity and hemodynamics variables. *IEEE transactions on biomedical circuits and systems*, *10*(2), 280–288.

Expert Panel on Detection, E. et al. (2001). Executive summary of the third report of the national cholesterol education program (ncep) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (adult treatment panel iii). *Jama*, *285*(19), 2486.

Fay, B. (1987). Critical social science: Liberation and its limits.

Feinstein, A. R. (1996). *Multivariable analysis: an introduction*. Yale University Press.

Ferrario, M., Chiodini, P., Chambless, L. E., Cesana, G., Vanuzzo, D., Panico, S., ... Giampaoli, S. (2005). Prediction of coronary events in a low incidence population. assessing accuracy of the cuore cohort study prediction equation. *International journal of epidemiology*, *34*(2), 413–421.

Framingham Heart Study. (2017, December). *Web tools based on Framingham heart study* [Internet web page]. Framingham Heart Study. Retrieved 2017, from `http://www.framinghamheartstudy.org`

Gill, E. & Mangin, D. (2011). Can general practitioners provide effective cardiovascular disease (cvd) prevention? dreams and realities of cvd prevention. *The New Zealand Medical Journal (Online)*, *124*(1328).

Given, L. M. (2008). *The sage encyclopedia of qualitative research methods.* Los Angeles, Calif. : Sage Publications, [2008].

Goff, D. C., Lloyd-Jones, D. M., Bennett, G., Coady, S., D'Agostino, R. B., Gibbons, R., ... others (2014). 2013 acc/aha guideline on the assessment of cardiovascular risk. *Circulation*, *129*(25 suppl 2), S49–S73.

Grambsch, P. M. & O'Brien, P. C. (1991). The effects of transformations and preliminary tests for non-linearity in regression. *Statistics in Medicine*, *10*(5), 697–709.

Green, B. B., Cook, A. J., Ralston, J. D., Fishman, P. A., Catz, S. L., Carlson, J., ... Thompson, R. S. (2008). Effectiveness of home blood pressure monitoring, web communication, and pharmacist care on hypertension control: a randomized controlled trial. *Jama*, *299*(24), 2857–2867.

Grundy, S. M., Cleeman, J. I., Daniels, S. R., Donato, K. A., Eckel, R. H., Franklin, B. A., ... others (2005). Diagnosis and management of the metabolic syndrome: an American Heart Association/National Heart, Lung, and Blood Institute scientific statement. *Circulation*, *112*(17), 2735–2752.

Guba, E. G. (1990). *The paradigm dialog.* Newbury Park, Calif. : Sage Publications, [1990].

Hachesu, P. R., Ahmadi, M., Alizadeh, S. & Sadoughi, F. (2013). Use of data mining techniques to determine and predict length of stay of cardiac patients. *Healthcare informatics research*, *19*(2), 121–129.

Hammermeister, K., DeRouen, T. A. & Dodge, H. T. (1979). Variables predictive of survival in patients with coronary disease. selection by univariate and multivariate analyses from the clinical, electrocardiographic, exercise, arteriographic, and quantitative angiographic evaluations. *Circulation*, *59*(3), 421–430.

Han, K. H., Park, K. C., Kim, M. J., Kim, Y. S. & Chun, H. (2017). Association between heart rate variability and 10-year atherosclerotic cardiovascular disease risk score. *Atherosclerosis*, *263*, e190–e191.

Hanley, J. A. & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, *143*(1), 29–36.

Harrell, F. (2013). Regression modeling strategies. *as implemented in R package 'rms' version*, *3*(3).

Harrell, J. F., Lee, K. L., Matchar, D. B. & Reichert, T. A. (1985). Regression models

for prognostic prediction: advantages, problems, and suggested solutions. *Cancer treatment reports*, *69*(10), 1071–1077.

Harriss, L. R., English, D. R., Powles, J., Giles, G. G., Tonkin, A. M., Hodge, A. M., ... O'dea, K. (2007). Dietary patterns and cardiovascular mortality in the Melbourne collaborative cohort study. *The American journal of clinical nutrition*, *86*(1), 221–229.

Hasuo, Y., Ueda, K., Kiyohara, Y., Wada, J., Kawano, H., Kato, I., ... Fujishima, M. (1989). Accuracy of diagnosis on death certificates for underlying causes of death in a long-term autopsy-based population study in Hisayama, Japan; with special reference to cardiovascular diseases. *Journal of clinical epidemiology*, *42*(6), 577–584.

Hay, D. R. (2004). *Cardiovascular disease in New Zealand, 2004: A summary of recent statistical information*. National Heart Foundation of New Zealand.

He, Y. & Zaslavsky, A. M. (2012). Diagnosing imputation models by applying target analyses to posterior replicates of completed data. *Statistics in medicine*, *31*(1), 1–18.

Heart Foundation. (2017, November). *General heart statistics in New Zealand* [Internet web page]. Heart Foundation. Retrieved 2017, from `https://www.heartfoundation.org.nz/statistics`

Heidenreich, P. A., Trogdon, J. G., Khavjou, O. A., Butler, J., Dracup, K., Ezekowitz, M. D., ... others (2011). Forecasting the future of cardiovascular disease in the united states: a policy statement from the American Heart Association. *Circulation*, *123*(8), 933–944.

Hermansson, J. & Kahan, T. (2018, 01 Feb). Systematic review of validity assessments of Framingham risk score results in health economic modelling of lipid-modifying therapies in Europe. *PharmacoEconomics*, *36*(2), 205–213. doi: 10.1007/s40273-017-0578-1

Hervás, R., Fontecha, J., Ausín, D., Castanedo, F., Bravo, J. & López-de Ipiña, D. (2013). Mobile monitoring and reasoning methods to prevent cardiovascular diseases. *Sensors*, *13*(5), 6524–6541.

Hippisley-Cox, J., Coupland, C., Vinogradova, Y., Robson, J. & Brindle, P. (2008). Performance of the qrisk cardiovascular risk prediction algorithm in an independent uk sample of patients from general practice: a validation study. *Heart*, *94*(1), 34–39.

Hippisley-Cox, J., Coupland, C., Vinogradova, Y., Robson, J., May, M. & Brindle, P. (2007). Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *Bmj*, *335*(7611), 136.

Horton, N. J. & Kleinman, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, *61*(1), 79–90.

Hua, X., McDermott, R., Lung, T., Wenitong, M., Tran-Duy, A., Li, M. & Clarke, P. (2017). Validation and recalibration of the Framingham cardiovascular disease risk models in an Australian indigenous cohort. *European journal of preventive*

cardiology, *24*(15), 1660–1669.

Hubert, H. B., Feinleib, M., McNamara, P. M. & Castelli, W. P. (1983). Obesity as an independent risk factor for cardiovascular disease: a 26-year follow-up of participants in the Framingham heart study. *Circulation*, *67*(5), 968–977.

Hurling, R., Catt, M., De Boni, M., Fairley, B. W., Hurst, T., Murray, P., ... Sodhi, J. S. (2007). Using internet and mobile phone technology to deliver an automated physical activity program: randomized controlled trial. *Journal of medical Internet research*, *9*(2).

Iasonos, A., Schrag, D., Raj, G. V. & Panageas, K. S. (2008). How to build and interpret a nomogram for cancer prognosis. *Journal of clinical oncology*, *26*(8), 1364–1370.

Jackson, R. (2000). Updated New Zealand cardiovascular disease risk-benefit prediction guide. *BMJ: British Medical Journal*, *320*(7236), 709.

Jee, S. H., Suh, I., Kim, I. S. & Appel, L. J. (1999). Smoking and atherosclerotic cardiovascular disease in men with low levels of serum cholesterol: the Korea Medical Insurance Corporation study. *Jama*, *282*(22), 2149–2155.

Jilani, T. A., Yasin, H., Yasin, M. & Ardil, C. (2009). Acute coronary syndrome prediction using data mining techniques-an application. *World Academy of Science, Engineering and Technology*, *59*(4), 295–299.

Jin, Z., Oresko, J., Huang, S. & Cheng, A. C. (2009). Hearttogo: a personalized medicine technology for cardiovascular disease prevention and detection. In *Life science systems and applications workshop, 2009. lissa 2009. ieee/nih* (pp. 80–83).

Jinks, R. C. (2012). *Sample size for multivariable prognostic models* (Unpublished doctoral dissertation). UCL (University College London).

Kannel, W. B., Abbott, R. D., Savage, D. D. & McNamara, P. M. (1982). Epidemiologic features of chronic atrial fibrillation: the Framingham study. *New England Journal of Medicine*, *306*(17), 1018–1022.

Kannel, W. B., Dawber, T. R., Kagan, A., Revotskie, N. & Stokes, J. (1961). Factors of risk in the development of coronary heart disease—six-year follow-up experience the Framingham study. *Annals of internal medicine*, *55*(1), 33–50.

Kannel, W. B., Feinleib, M., McNamara, P. M., Garrison, R. J. & Castelli, W. P. (1979). An investigation of coronary heart disease in families: the Framingham offspring study. *American journal of epidemiology*, *110*(3), 281–290.

Kannel, W. B., McGee, D. & Gordon, T. (1976). A general cardiovascular risk profile: the Framingham study. *American Journal of Cardiology*, *38*(1), 46–51.

Kano, Y. & Harada, A. (2000, 01 Mar). Stepwise variable selection in factor analysis. *Psychometrika*, *65*(1), 7–22. doi: 10.1007/BF02294182

Kaplan, E. L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, *53*(282), 457–481.

Kattan, M. W. (2003a). Comparison of cox regression with other methods for determining prediction models and nomograms. *The Journal of urology*, *170*(6), S6–S10.

Kattan, M. W. (2003b). Nomograms are superior to staging and risk grouping systems

for identifying high-risk patients: preoperative application in prostate cancer. *Current Opinion in Urology*, *13*(2), 111–116.

Katz, M. H. (2011). *Multivariable analysis: a practical guide for clinicians and public health researchers*. Cambridge university press.

Kendall, M. G. (1955). Rank correlation methods.

Kenward, M. G. (2013). The handling of missing data in clinical trials. *Clinical Investigation*, *3*(3), 241–250.

Kim, H., Ishag, M. I. M., Piao, M., Kwon, T. & Ryu, K. H. (2016). A data mining approach for cardiovascular disease diagnosis using heart rate variability and images of carotid arteries. *Symmetry*, *8*(6), 47.

Kim, J., Lee, J. & Lee, Y. (2015). Data-mining-based coronary heart disease risk prediction model using fuzzy logic and decision tree. *Healthcare informatics research*, *21*(3), 167–174.

Kleinbaum, D. G. & Klein, M. (2010). *Survival analysis* (Vol. 3). Springer.

Kleinbaum, D. G., Kupper, L. L., Nizam, A. & Rosenberg, E. S. (2013). *Applied regression analysis and other multivariable methods*. Cengage Learning.

Koh, H. C., Tan, G. et al. (2011). Data mining applications in healthcare. *Journal of healthcare information management*, *19*(2), 65.

Kuhfeld, W. F. (1990). The prinqual procedure. *SAS/STAT User's Guide*, *2*, 1265–1323.

Kumari, M. & Godara, S. (2011). Comparative study of data mining classification methods in cardiovascular disease prediction 1.

Le, H., Marchant, I., Subtil, F., Boissel, J. & Gueyffier, F. (2017). Validation of Framingham and score in 30560 patients with hypertension. *Journal of Hypertension*, *35*, e28.

Leavy, P. (2017). *Research design: Quantitative, qualitative, mixed methods, arts-based, and community-based participatory research approaches*. Guilford Publications.

Lee, E. T. & Wang, J. (2003). *Statistical methods for survival data analysis* (Vol. 476). John Wiley & Sons.

Lee, W., Yoon, H. & Park, K. (2016). Smart ecg monitoring patch with built-in r-peak detection for long-term hrv analysis. *Annals of biomedical engineering*, *44*(7), 2292–2301.

Lin, C.-W., Yang, Y.-T. C., Wang, J.-S. & Yang, Y.-C. (2012). A wearable sensor module with a neural-network-based activity classification algorithm for daily energy expenditure estimation. *IEEE Transactions on Information Technology in Biomedicine*, *16*(5), 991–998.

Liu, M., Wu, B., Wang, W.-Z., Lee, L.-M., Zhang, S.-H. & Kong, L.-Z. (2007). Stroke in china: epidemiology, prevention, and management strategies. *The Lancet Neurology*, *6*(5), 456–464.

Liu, Y.-M., Chen, S. L.-S., Yen, A. M.-F. & Chen, H.-H. (2013). Individual risk prediction model for incident cardiovascular disease: A bayesian clinical reasoning approach. *International journal of cardiology*, *167*(5), 2008–2012.

Lloyd-Jones, D. M., Wilson, P. W., Larson, M. G., Beiser, A., Leip, E. P., D'Agostino, R. B. & Levy, D. (2004). Framingham risk score and prediction of lifetime risk for coronary heart disease. *The American journal of cardiology*, *94*(1), 20–24.

Lopez, A. D., Mathers, C. D., Ezzati, M., Jamison, D. T. & Murray, C. J. (2006). Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. *The Lancet*, *367*(9524), 1747–1757.

Lymberis, A. (2003). Smart wearable systems for personalised health management: current r&d and future challenges. In *Engineering in medicine and biology society, 2003. proceedings of the 25th annual international conference of the ieee* (Vol. 4, pp. 3716–3719).

Ma, J., Li, H., Giovannucci, E., Mucci, L., Qiu, W., Nguyen, P. L., ... Stampfer, M. J. (2008). Prediagnostic body-mass index, plasma c-peptide concentration, and prostate cancer-specific mortality in men with prostate cancer: a long-term survival analysis. *The lancet oncology*, *9*(11), 1039–1047.

Mahmood, S. S., Levy, D., Vasan, R. S. & Wang, T. J. (2014). The Framingham heart study and the epidemiology of cardiovascular disease: a historical perspective. *The Lancet*, *383*(9921), 999–1008.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother. Rep.*, *50*, 163–170.

McGill, H. C., McMahan, C. A. & Gidding, S. S. (2008). Preventing heart disease in the 21st century. *Circulation*, *117*(9), 1216–1227.

Mehta, D. (2005). *British national formulary* (Vol. 49). Pharmaceutical Press.

Melillo, P., Izzo, R., Orrico, A., Scala, P., Attanasio, M., Mirra, M., ... Pecchia, L. (2015). Automatic prediction of cardiovascular and cerebrovascular events using heart rate variability analysis. *PloS one*, *10*(3), e0118504.

Mendis, S. (2010). The contribution of the Framingham heart study to the prevention of cardiovascular disease: a global perspective. *Progress in cardiovascular diseases*, *53*(1), 10–14.

Mendis, S., Puska, P., Norrving, B. et al. (2011). *Global atlas on cardiovascular disease prevention and control.* World Health Organization.

Merriam, S. B. & Tisdell, E. J. (2016). *Qualitative research : a guide to design and implementation.* San Francisco, CA : Jossey-Bass, [2016].

Mertens, D. M. (2014). *Research and evaluation in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods*. Sage publications.

Michard, F. (2017). A sneak peek into digital innovations and wearable sensors for cardiac monitoring. *Journal of clinical monitoring and computing*, *31*(2), 253–259.

Milani, R. V. & Franklin, N. C. (2017). The role of technology in healthy living medicine. *Progress in Cardiovascular Diseases*.

Milenković, A., Otto, C. & Jovanov, E. (2006). Wireless sensor networks for personal health monitoring: Issues and an implementation. *Computer communications*, *29*(13), 2521–2533.

Miller Jr, R. G. (2011). *Survival analysis* (Vol. 66). John Wiley & Sons.

Ministry of Health. (2018). *Cardiovascular disease risk assessment and management for primary care.* Wellington: Ministry of Health.

Moons, K. G., Altman, D. G., Vergouwe, Y. & Royston, P. (2009). Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *Bmj*, *338*, b606.

Moons, K. G., de Groot, J. A., Bouwmeester, W., Vergouwe, Y., Mallett, S., Altman, D. G., ... Collins, G. S. (2014). Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the charms checklist. *PLoS medicine*, *11*(10), e1001744.

Moons, K. G., Kengne, A. P., Woodward, M., Royston, P., Vergouwe, Y., Altman, D. G. & Grobbee, D. E. (2012). Risk prediction models: I. development, internal validation, and assessing the incremental value of a new (bio) marker. *Heart*, *98*(9), 683–690.

Morse, J. M. & Niehaus, L. (2009). Walnut Creek, Calif. : Left Coast Press, [2009].

Mostafa, S., Davies, M., Webb, D., Gray, L., Srinivasan, B., Jarvis, J. & Khunti, K. (2010). The potential impact of using glycated haemoglobin as the preferred diagnostic tool for detecting type 2 diabetes mellitus. *Diabetic Medicine*, *27*(7), 762–769.

Mozaffarian, D., Benjamin, E. J., Go, A. S., Arnett, D. K., Blaha, M. J., Cushman, M., ... others (2015). Heart disease and stroke statistics—2015 update: a report from the american heart association. *Circulation*, *131*(4), e29–e322.

Murukesan, L., Murugappan, M., Iqbal, M. & Saravanan, K. (2014). Machine learning approach for sudden cardiac arrest prediction based on optimal heart rate variability features. *Journal of Medical Imaging and Health Informatics*, *4*(4), 521–532.

National Heart, Lung, and Blood Institute. (2018, Januray). *National heart, lung, and blood institute.* National Heart, Lung, and Blood Institute. Retrieved 2018, from `https://www.nhlbi.nih.gov/`

Neuman, L. W. (2002). Social research methods: Qualitative and quantitative approaches.

Odden, M. C., Shlipak, M. G., Whitson, H. E., Katz, R., Kearney, P. M., defilippi, C., ... Newman, A. B. (2014). Risk factors for cardiovascular disease across the spectrum of older age: The cardiovascular health study. *Atherosclerosis*, *237*, 336 - 342.

Oresko, J. J., Jin, Z., Cheng, J., Huang, S., Sun, Y., Duschl, H. & Cheng, A. C. (2010). A wearable smartphone-based platform for real-time cardiovascular disease detection via electrocardiogram processing. *IEEE Transactions on Information Technology in Biomedicine*, *14*(3), 734–740.

Pantelopoulos, A. & Bourbakis, N. G. (2010). A survey on wearable sensor-based systems for health monitoring and prognosis. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(1), 1–12.

Parikh, N. I., Hwang, S.-J., Larson, M. G., Cupples, L. A., Fox, C. S., Manders, E. S., ... O'Donnell, C. J. (2007). Parental occurrence of premature cardiovascular disease predicts increased coronary artery and abdominal aortic calcification in the Framingham offspring and third generation cohorts. *Circulation*, *116*(13), 1473–1481.

Parmar, P., Krishnamurthi, R., Ikram, M. A., Hofman, A., Mirza, S. S., Varakin, Y., ... others (2015). The stroke riskometertm app: Validation of a data collection tool and stroke risk predictor. *International Journal of Stroke*, *10*(2), 231–244.

Patel, S., Park, H., Bonato, P., Chan, L. & Rodgers, M. (2012). A review of wearable sensors and systems with application in rehabilitation. *Journal of neuroengineering and rehabilitation*, *9*(1), 21.

Patrick, K., Raab, F., Adams, M. A., Dillon, L., Zabinski, M., Rock, C. L., ... Norman, G. J. (2009). A text message–based intervention for weight loss: randomized controlled trial. *Journal of medical Internet research*, *11*(1).

Pencina, M. J., D'agostino, R. B., Larson, M. G., Massaro, J. M. & Vasan, R. S. (2009). Predicting the 30-year risk of cardiovascular disease: the Framingham heart study. *Circulation*, *119*(24), 3078–3084.

Phillips, D. C. & Burbules, N. C. (2000). *Postpositivism and educational research*. Rowman & Littlefield.

Rorty, R. (1990). Pragmatism as anti-representationalism. *JP Murphy, Pragmatism: From Peirce to Davison*, 1–6.

Royston, P. & Altman, D. G. (2013). External validation of a cox prognostic model: principles and methods. *BMC medical research methodology*, *13*(1), 33.

Royston, P. et al. (2004). Multiple imputation of missing values. *Stata journal*, *4*(3), 227–41.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American statistical Association*, *91*(434), 473–489.

Savin-Baden, M. & Major, C. H. (2013). *Qualitative research : the essential guide to theory and practice.* London : Routledge, Taylor Francis Group, 2013.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, *69*(1), 239–241.

Schumacher, M., Holländer, N. & Sauerbrei, W. (1997). Resampling and cross-validation techniques: a tool to reduce bias caused by model building? *Statistics in medicine*, *16*(24), 2813–2827.

Smith, J. K. (1983). Quantitative versus qualitative research: An attempt to clarify the issue. *Educational researcher*, *12*(3), 6–13.

Somers, R. H. (1962). A new asymmetric measure of association for ordinal variables. *American sociological review*, 799–811.

Splansky, G. L., Corey, D., Yang, Q., Atwood, L. D., Cupples, L. A., Benjamin, E. J., ... others (2007). The third generation cohort of the national heart, lung, and blood institute's Framingham heart study: design, recruitment, and initial examination. *American journal of epidemiology*, *165*(11), 1328–1335.

Statistics New Zealand. (2012). *Demographic trends 2010.* Wellington: Statistics New Zealand ISSN.

Statistics New Zealand and Ministry of Pacific Island Affairs. (2011). *Health and pacific peoples in New Zealand.* Wellington: Statistics New Zealand and Ministry of Pacific Island Affairs.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., ... Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological

and clinical research: potential and pitfalls. *Bmj*, *338*, b2393.

Tartakovsky, A., Nikiforov, I. & Basseville, M. (2014). *Sequential analysis: Hypothesis testing and changepoint detection*. Chapman and Hall/CRC.

Tashakkori, A. & Teddlie, C. (2010). *Sage handbook of mixed methods in social behavioral research.* Los Angeles : SAGE Publications, [2010].

Team, R. C. et al. (2013). R: A language and environment for statistical computing.

Unnikrishnan, P., Kumar, D. K., Poosapadi Arjunan, S., Kumar, H., Mitchell, P. & Kawasaki, R. (2016). Development of health parameter model for risk prediction of cvd using svm. *Computational and mathematical methods in medicine*, *2016*.

Vaanathi, S. (2017). Cardiovascular disease prediction using fuzzy logic expert system. *IUP Journal of Computer Sciences*, *11*(3).

Van Buuren, S. (2012). *Flexible imputation of missing data*. CRC press.

Weiner, D. E., Tighiouart, H., Amin, M. G., Stark, P. C., MacLeod, B., Griffith, J. L., ... Sarnak, M. J. (2004). Chronic kidney disease as a risk factor for cardiovascular disease and all-cause mortality: a pooled analysis of community-based studies. *Journal of the American Society of Nephrology*, *15*(5), 1307–1315.

Wells, S., Kerr, A., Eadie, S., Wiltshire, C. & Jackson, R. (2010). 'your heart forecast': a new approach for describing and communicating cardiovascular risk? *Heart*, *96*(9), 708–713.

Wells, S., Riddell, T., Kerr, A., Pylypchuk, R., Chelimo, C., Marshall, R., ... others (2017). Cohort profile: the predict cardiovascular disease cohort in New Zealand primary care (predict-cvd 19). *International journal of epidemiology*, *46*(1), 22–22.

Wessler, B. S., Kramer, W., Cangelosi, M., Raman, G., Lutz, J. S. & Kent, D. M. (2015). Clinical prediction models for cardiovascular disease: tufts predictive analytics and comparative effectiveness clinical prediction model database. *Circulation: Cardiovascular Quality and Outcomes*, *8*(4), 368–375.

White, I. R., Royston, P. & Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, *30*(4), 377–399.

Whitehead, J. (1993). Sample size calculations for ordered categorical data. *Statistics in medicine*, *12*(24), 2257–2271.

Whittaker, R., McRobbie, H., Bullen, C., Borland, R., Rodgers, A. & Gu, Y. (2012). Mobile phone-based interventions for smoking cessation. *The Cochrane Library*.

Wilson, P. W., D'agostino, R. B., Sullivan, L., Parise, H. & Kannel, W. B. (2002). Overweight and obesity as determinants of cardiovascular risk: the Framingham experience. *Archives of internal medicine*, *162*(16), 1867–1872.

Wilson, P. W., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H. & Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation*, *97*(18), 1837–1847.

Woodward, M., Brindle, P. & Tunstall-Pedoe, H. (2006). Adding social deprivation and family history to cardiovascular risk assessment-the ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC). *Heart*.

World Health Organisation and others. (n.d.). *mhealth: new horizons for health through mobile technologies: based on the findings of the second survey on ehealth. geneva: Who; 2011.*

Yu, J., Dai, L., Zhao, Q., Liu, X., Chen, S., Wang, A., ... Wu, S. (2017). Association of cumulative exposure to resting heart rate with risk of stroke in general population: The Kailuan cohort study. *Journal of Stroke and Cerebrovascular Diseases*, *26*(11), 2501–2509.

Zhang, X.-F., Attia, J., D'Este, C., Yu, X.-H. & Wu, X.-G. (2005). A risk score predicted coronary heart disease and stroke in a Chinese cohort. *Journal of clinical epidemiology*, *58*(9), 951–958.

Appendix A

Abbreviations

CVD Cardiovascular Disease

CHD Coronary Heart Disease

CAD Coronary Artery Diseases

EHRs Electronic Health Records

SHHEC Scottish Heart Health Extended Cohort

PROCAM Prospective Cardiovascular Münster

cumRHR cumulative exposure to Resting Heart Rate

ASCVD Atherosclerotic Cardiovascular Disease

SCA Sudden Cardiac Arrest

HDL High Density Lipoprotein

LDL Low Density Lipoprotein

SBP Systolic Blood Pressure

HF Heart Failure

IHF Incident Heart Failure

CPMs Clinical Prediction Models

T2DM Type 2 Diabetes Mellitus

CKD Chronic Kidney Disease

SVM Support Vector Machines

ANN Artificial Neural Network

CART Classification and Regression Tree

KNHANES-VI Korean National Health and Nutrition Examination Survey VI

PNN Probabilistic Neural Net-work

AI Artificial Intelligence

BMI Body Mass Index

CRP C-reactive Protein

HRV Heart Rate Variability

KCIS Keelung Community-based Integrated Screening

NHLBI National Heart, Lung and Blood Institute

JBS Joint British Societies

BNF British National Formulary

WHO World Health Organisation

GOe Global Observatory for eHealth

PDAs Personal Digital Assistants

MMM My Meal Mate

ECG Electrocardiography

RBFN Radial Basis Function Network

GRNN Generalised Regression Neural Network

CHARMS CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies

FHS Framingham Heart Study

HR Hazard Ratios

MCAR Missing Completely At Random

MAR Missing At Random

IM Informative Missing

MGV Maximum Generalized Variance

PLS Partial Least Squares

Lasso Least absolute shrinkage and selection operator

PH Proportional Hazards

GOF Goodness-of-fit

ROC Receiver Operating Characteristic

Appendix B

Ethics Approval

AUTEC Secretariat

Auckland University of Technology
D-88, WU406 Level 4 WU Building City Campus
T: +64 9 921 9999 ext. 8316
E: ethics@aut.ac.nz
www.aut.ac.nz/researchethics

14 November 2017

Farhaan Mirza
Faculty of Design and Creative Technologies

Dear Farhaan

Re Ethics Application: **17/385 Early detection and self-management of cardiovascular disease using Artificial Intelligence-based Model**

Thank you for providing evidence as requested, which satisfies the points raised by the Auckland University of Technology Ethics Committee (AUTEC).

Your ethics application has been approved for three years until 14 November 2020.

Standard Conditions of Approval

1. A progress report is due annually on the anniversary of the approval date, using form EA2, which is available online through http://www.aut.ac.nz/researchethics.
2. A final report is due at the expiration of the approval period, or, upon completion of project, using form EA3, which is available online through http://www.aut.ac.nz/researchethics.
3. Any amendments to the project must be approved by AUTEC prior to being implemented. Amendments can be requested using the EA2 form: http://www.aut.ac.nz/researchethics.
4. Any serious or unexpected adverse events must be reported to AUTEC Secretariat as a matter of priority.
5. Any unforeseen events that might affect continued ethical acceptability of the project should also be reported to the AUTEC Secretariat as a matter of priority.

Please quote the application number and title on all future correspondence related to this project.

AUTEC grants ethical approval only. If you require management approval for access for your research from another institution or organisation then you are responsible for obtaining it. You are reminded that it is your responsibility to ensure that the spelling and grammar of documents being provided to participants or external organisations is of a high standard.

For any enquiries, please contact ethics@aut.ac.nz

Yours sincerely,

Kate O'Connor
Executive Manager
Auckland University of Technology Ethics Committee

Cc: nfx1464@autuni.ac.nz; Hamid Gholam Hosseini; mirza.baig@aut.ac.nz

Appendix C

Complete Data Columns

ORDERS IN DOCUMNETS	FACTORS NAME
MF3	SEX
MF4	RELATIVE WEIGHT
MF5	EDUCATION
MF6	COUNTRY OF BIRTH
MF7	ITALIAN ANCESTRY
MF8	HISTORY OF ACUTE INFECTIONS
MF9	HISTORY OF RHEUMATIC FEVER
MF10	HISTORY OF ALLERGY OR ASTHMA
MF11	HISTORY OF CHRONIC ARTHRITIS AND RHEUMATISM
MF12	HISTORY OF THYROID DISEASE
MF13	HISTORY OF HYPERTENSION
MF14	HISTORY OF ENLARGED HEART
MF15	HISTORY OF NERVOUS HEART
MF16	HISTORY OF PERICARDITIS
MF17	HISTORY OF SUBACUTE ENDOCARDITIS
MF18	HISTORY OF OTHER CARDIOVASCULAR DISEASE
MF19	HISTORY OF NON-CARDIOVASCULAR DISEASE
MF20	NUMBER OF PREGNANCIES
MF21	TOXEMIA WITH ONE OR MORE PREGNANCIES
MF22	HISTORY OF DRUGS TAKEN
MF23	INCREASE IN CHEST CIRCUMFERENCE
MF24	EYE EXAMINATION: EXOPHTHALMOS, ARCUS SENILIS, XANTHELASMA
MF25	EYE EXAMINATION: RETINA
MF26	CHEST EXAMINATION: DEFORMITY
MF27	CHEST EXAMINATION: APEX IMPULSE OUTSIDE MIDCLAVICULAR, LINE
MF28	CHEST EXAMINATION: RALES
MF29	CHEST EXAMINATION: IRREGULAR CARDIAC RHYTHM
MF30	X-RAY: GENERALIZED HYPERTROPHY
MF31	X-RAY: LEFT VENTRICULAR HYPERTROPHY
MF32	X-RAY: OTHER HYPERTROPHY (AH, RVH)
MF33	X-RAY: ABNORMAL CONTOUR OTHER THAN HYPERTROPHY
MF34	X-RAY: ABNORMAL GREAT VESSELS
MF35	X-RAY: ABNORMAL POSITION
MF36	X-RAY: ABNORMAL CALCIFICATION EXCEPT IN AORTA
MF37	X-RAY: OTHER CARDIOVASCULAR ABNORMALITIES
MF38	X-RAY: NON-CARDIOVASCULAR ABNORMALITIES
MF39	ECG: GENERAL DIAGNOSIS
BMI1	BODY MASS INDEX

ORDERS IN DOCUMNETS	FACTORS NAME
MF40	ECG: ATRIOVENTRICULAR BLOCK
MF41	ECG: AURICULAR FIBRILLATION OR FLUTTER
MF42	ECG: PREMATURE BEATS
MF43	ECG: MYOCARDIAL INFARCTION
MF44	ECG: VENTRICULAR HYPERTROPHY OR STRAIN
MF45	ECG: NON-SPECIFICALLY ABNORMAL
MF46	ECG: OTHER ABNORMALITY
MF47	FINAL DIAGNOSTIC IMPRESSION: CONGESTIVE HEART FAILURE
MF48	URINALYSIS: SUGAR
MF49	URINALYSIS: ALBUMIN
MF50	FINAL DIAGNOSTIC IMPRESSION: DIABETES
MF51	FINAL DIAGNOSTIC IMPRESSION: ANEMIA
MF52	FINAL DIAGNOSTIC IMPRESSION: THYROID DISEASE
MF53	PHLEBITIS
MF54	FINAL DIAGNOSTIC IMPRESSION: OTHER PERIPHERAL VASCULAR DISEASE
MF55	FINAL DIAGNOSTIC IMPRESSION: LIVER DISEASE
MF56	BLOOD PRESSURE: ADMISSION, SYSTOLIC
MF57	BLOOD PRESSURE: ADMISSION, DIASTOLIC
MF58	BLOOD PRESSURE: FIRST EXAMINER, SYSTOLIC
MF59	BLOOD PRESSURE: FIRST EXAMINER, DIASTOLIC
MF60	BLOOD PRESSURE: SECOND EXAMINER, SYSTOLIC
MF61	BLOOD PRESSURE: SECOND EXAMINER, DIASTOLIC
MF62	BLOOD ANALYSIS: SERUM CHOLESTEROL
MF63	BLOOD ANALYSIS: HEMOGLOBIN
MF64	BLOOD ANALYSIS: PHOSPHOLIPID
MF65	BLOOD ANALYSIS: SUGAR
MF66	BLOOD ANALYSIS: URIC ACID
MF67	HEIGHT: FULL INCHES
MF68	HEIGHT: IN EXCESS OF FULL INCHES
MF69	WEIGHT
MF70	HISTORY FORMS USED AS SOURCE FOR SMOKING AND DRINKING DATA
MF71	TOBACCO USED "NOW" OR "EVER"
MF72	DURATION OF TOBACCO USE (WHEN SMOKING "NOW")
MF73	PERIOD OF LAPSE IN TOBACCO USE (WHEN NOT SMOKING "NOW")
MF74	CIGARETTES SMOKED PER DAY
MF75	CIGARS SMOKED PER DAY
MF76	PIPES SMOKED PER DAY
MF77	OTHER TOBACCO USED

Appendix D

Code

```
covariates <- c("age", "sex", "bmi", "hyp", "honh", "hop", "hooc",
          "pb", "hoarb", "horf", "hoaoa", "hotd", "hose",   "bps",
          "bpd", "cgrpd", "cgpd", "ppd", "pr", "dia")

forward_formulas <- sapply(covariates,
          function(x) as.formula(paste('Surv(cvddate, cvd)~', x)))

forward_models <- lapply( forward_formulas,
          function(x){coxph(x, data = FOCExam1NoMissingData)})

results <- lapply(forward_models,
          function(x){
          x <- summary(x)
          p.value<-signif(x$wald["pvalue"], digits=2)
          wald.test<-signif(x$wald["test"], digits=2)
          beta<-signif(x$coef[1], digits=2);#coeficient beta
          HR <-signif(x$coef[2], digits=2);#exp(beta)
          HR.confint.lower <- signif(x$conf.int[,"lower .95"], 2)
          HR.confint.upper <- signif(x$conf.int[,"upper .95"],2)
          HR <- paste0(HR, " (", HR.confint.lower, "-", HR.confint.upper,")")
          res<-c(beta, HR, wald.test, p.value)
          names(res)<-c("beta", "HR (95% CI for HR)", "wald.test", "p.value")
          return(res) })

res <- t(as.data.frame(results, check.names = TRUE))
as.data.frame(res)
```

Code Snippet D.1: R Code of Selecting Variables using Forward Selection

```
#processCsv.py
import os
import logging
import hashlib
import csv

def file_name(file_dir):
    for root, dirs, files in os.walk(file_dir):
        print(root) # current directory
        print(dirs) # sub directory in current directory
        print(files) # files in current directory

    path_out = 'result.txt'
    file_out = open(path_out,'w')

    target = getPIDArray()
    columns = getColumns()
    col_names = []
    for index, column in enumerate(columns):
        print(index, column)
        col_names.append(column)

    print("\n csv files number is ",len(files))
    for i in range (len(files)):
        print("\n ",i ,files[i])
        file_out.write(" ------------\n ")
        file_out.write(files[i]+" :\n ")
        searchInCsvFile(files[i],target,columns,col_names)
    file_out.close()
def getPIDArray():
    fileName = "PID.csv"
    csvFile = open(fileName, "r")
    reader = csv.reader(csvFile)
    pid=[]
    for item in reader:
        patientID = item[0]
        pid.append(patientID)
    return pid
def getColumns():
    fileName = "Columns.csv"
    csvFile = open(fileName, "r")
    reader = csv.reader(csvFile)
    columns={}
    for item in reader:
        filename = item[0]
        columns[item[0]]=[]
        for i in range (1,len(item)):
            columns[item[0]].append(item[i])
        print("columns name is: ",columns)
    return columns
def searchInCsvFile(file, targetPID, columns ,col_names):
    print("\n enter searchInCsvFile ")
    if not file in col_names:
        print("\n file is not concerned ")
```

```
55          return
56      path_out = "res/"+file+'_res.csv'
57      outputFile = open(path_out,'w')
58      for col in columns[file]:
59          outputFile.write(col+",")
60      outputFile.write("\n")
61
62      fileName = "csv/"+file
63      csvFile = open(fileName, "r")
64      reader = csv.reader(csvFile)
65      count = 0
66      pid_col = -1
67      for item in reader:
68          count += 1
69          if count == 1:
70              title = item
71              col_num = len(item)
72              for i in range (col_num):
73                  if item[i] == "PID":
74                      print("\n PID column is: ",i)
75                      pid_col = i
76          elif count>1:
77              if pid_col>0:
78                  if item[pid_col] in targetPID:
79                      value = {}
80                      for index in range (col_num):
81                          value[title[index]] = item[index]
82                      concern_col_num = len(columns[file])
83                      for j in range (concern_col_num):       outputFile.write(value[columns[file][j]]+",")
84                      outputFile.write("\n")
85      outputFile.close()
86  file_name("csv")
```

Code Snippet D.2: Python Code for Data Extraction

Appendix E

Research Outputs from Thesis

- **The general CVD risk prediction model present in Chapter 6.**